STA261 (Summer 2024) - Assignment 0

Solutions

These problems are meant to refresh/flex your STA257 skills (and your calculus skills). They are *not* to be handed in. Problems marked with stars (*) are results that will be used later in our course.

---

1. (a) Let $X \sim \mathcal{N}\left(0, \sigma^2\right)$. Show that $\mathbb{E}\left[X^{2k+1}\right] = 0$ for any $k \in \mathbb{N}$.

Since the integrand of $\mathbb{E}\left[X^{2k+1}\right]$ is odd, the result follows immediately. To spell out the details a bit,

$$\begin{aligned}
&\mathbb{E}\left[X^{2k+1}\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x^{2k+1} \cdot e^{-x^2/2\sigma^2} \, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(\int_{-\infty}^{0} x^{2k+1} \cdot e^{-x^2/2\sigma^2} \, \mathrm{d}x + \int_{0}^{\infty} x^{2k+1} \cdot e^{-x^2/2\sigma^2} \, \mathrm{d}x \right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \left(- \int_{0}^{\infty} u^{2k+1} \cdot e^{-u^2/2\sigma^2} \, \mathrm{d}u + \int_{0}^{\infty} x^{2k+1} \cdot e^{-x^2/2\sigma^2} \, \mathrm{d}x \right) \quad \text{Substituting in } u(x) = -x \text{ into the first integral and rearranging} \\
&= 0.
\end{aligned}$$

(b) Go a bit further and show that this is true for *any* continuous distribution which is symmetric about zero (i.e., its pdf satisfies $f_X(x) = f_X(-x)$ for any $x \in \mathbb{R}$), provided all of its moments are finite of course. In other words, if a distribution is symmetric about zero, then *all* of its odd moments must vanish. Can you generalize this result to distributions symmetric about an arbitrary point x_0 (i.e., those whose pdf satisfies $f_X(x_0 + x) = f_X(x_0 - x)$ for any $x \in \mathbb{R}$)?

Precisely the same reasoning as above applies, just with $e^{-x^2/2\sigma^2}/\sqrt{2\pi}\sigma$ replaced by the more general pdf $f_X(x)$. The generalization is simply $\mathbb{E}\left[(X - x_0)^{2k+1}\right] = 0$, because

$$\begin{aligned}
&\mathbb{E}\left[(X - x_0)^{2k+1}\right] \\
&= \int_{-\infty}^{\infty} (x - x_0)^{2k+1} \cdot f_X(x) \, \mathrm{d}x \\
&= \int_{-\infty}^{\infty} u^{2k+1} \cdot f_X(x_0 + u) \, \mathrm{d}u \quad \text{Substituting in } u(x) = x - x_0 \\
&= \int_{-\infty}^{0} u^{2k+1} \cdot f_X(x_0 + u) \, \mathrm{d}u + \int_{0}^{\infty} u^{2k+1} \cdot f_X(x_0 + u) \, \mathrm{d}u \\
&= - \int_{0}^{\infty} v^{2k+1} \cdot f_X(x_0 - v) \, \mathrm{d}v + \int_{0}^{\infty} u^{2k+1} \cdot f_X(x_0 + u) \, \mathrm{d}u \quad \text{Substituting in } v(u) = -u \text{ into the first integral and rearranging}
\end{aligned}$$

$$= -\int_0^\infty v^{2k+1} \cdot f_X(x_0+v)\,\mathrm{d}v + \int_0^\infty u^{2k+1} \cdot f_X(x_0+u)\,\mathrm{d}u \qquad \text{Since } f(x_0-v) = f(x_0+v)$$
$$= 0.$$

2. For any two (possibly dependent) random variables with finite second moments, show that

$$\mathrm{Var}\,(X+Y) + \mathrm{Var}\,(X-Y) = 2(\mathrm{Var}\,(X) + \mathrm{Var}\,(Y)).$$

This falls out from applying the identity

$$\mathrm{Var}\,(aX+bY) = a^2 \cdot \mathrm{Var}\,(X) + b^2 \cdot \mathrm{Var}\,(Y) + 2ab \cdot \mathrm{Cov}\,(X,Y)$$

twice.

3. Let $X \sim \mathrm{Poisson}\,(\lambda)$ and let $h : \mathbb{N} \to \mathbb{R}$ be any function such that $\mathbb{E}\,[h(X)]$ is finite. Prove that $\mathbb{E}\,[\lambda \cdot h(X)] = \mathbb{E}\,[X \cdot h(X-1)]$.

We have that

$$\begin{aligned}
\mathbb{E}\,[\lambda \cdot h(X)] &= \sum_{j\geq 0} \frac{\lambda^j \cdot e^{-\lambda}}{j!} \cdot \lambda \cdot h(j) \\
&= \sum_{j\geq 0} \frac{\lambda^{j+1} \cdot e^{-\lambda}}{j!} \cdot h(j) \\
&= \sum_{k\geq 1} \frac{\lambda^k \cdot e^{-\lambda}}{(k-1)!} \cdot h(k-1) && \text{Substituting } k = j+1 \\
&= \sum_{k\geq 1} \frac{\lambda^k \cdot e^{-\lambda}}{k!} \cdot k \cdot h(k-1) \\
&= \sum_{k\geq 0} \frac{\lambda^k \cdot e^{-\lambda}}{k!} \cdot k \cdot h(k-1) && \text{Since the summand is } 0 \text{ when } k = 0 \\
&= \mathbb{E}\,[X \cdot h(X-1)]\,.
\end{aligned}$$

4. Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and let $g : \mathbb{R} \to \mathbb{R}$ be any differentiable function that's nice enough to satisfy $\mathbb{E}\,[|g'(X)|] < \infty$ and $\lim_{|x|\to\infty} g(x) \cdot e^{-(x-\mu)^2/2\sigma^2} = 0$. Prove that $\mathbb{E}\,[g(X) \cdot (X-\mu)] = \sigma^2 \cdot \mathbb{E}\,[g'(X)]$. This is called *Stein's lemma* (in fact the condition that $\lim_{|x|\to\infty} g(x) \cdot e^{-(x-\mu)^2/2\sigma^2} = 0$ is unnecessary, but proving that is a lot harder).

Using integration by parts with $u(x) = e^{-(x-\mu)^2/2\sigma^2}$ and $\mathrm{d}v = g'(x)\,\mathrm{d}x$, we get

$$\begin{aligned}
\sigma^2 \cdot \mathbb{E}\left[g'(X)\right] &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(x) \cdot e^{-(x-\mu)^2/2\sigma^2}\,\mathrm{d}x \\
&= \frac{\sigma}{\sqrt{2\pi}} \left(e^{-(x-\mu)^2/2\sigma^2} \cdot g(x) \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} g(x) \cdot (x-\mu) \cdot e^{-(x-\mu)^2/2\sigma^2}\,\mathrm{d}x \right) \\
&= \frac{\sigma}{\sqrt{2\pi}} \cdot e^{-(x-\mu)^2/2\sigma^2} \cdot g(x) \Big|_{-\infty}^{\infty} + \mathbb{E}\,[g(X) \cdot (X-\mu)]\,,
\end{aligned}$$

and the condition $\lim_{|x|\to\infty} g(x) \cdot e^{-(x-\mu)^2/2\sigma^2} = 0$ shows that the term on the left is 0.

⋆5. For any set of univariate random variables $X_1, X_2, \ldots, X_n$, the *order statistics* are the X_i's placed in ascending order, which are notated as $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. Thus the *sample minimum* $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and the *sample maximum* $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

In STA257, you may have learned that if $X_1, X_2, \ldots, X_n$ are an independent sample from a continuous distribution with pdf f_X and cdf F_X, then $f_{X_{(1)}}(x) = n \cdot f_X(x) \cdot (1 - F_X(x))^{n-1}$ and $f_{X_{(n)}}(x) = n \cdot f_X(x) \cdot F_X(x)^{n-1}$. Let's generalize those formulas by finding the pdf of $X_{(j)}$, for any $1 \le j \le n$.

(a) Let $h > 0$ be nice and small. Explain why

$$\begin{aligned} &\mathbb{P}\left(X_{(j)} \in [x, x+h]\right) \\ &= \mathbb{P}\left(\text{One of the } X_i\text{'s is in } [x, x+h] \text{ and exactly } j-1 \text{ of the others are } < x\right). \end{aligned}$$

Because the X_i's are continuous, when h is small enough there's at most one of them in the interval $[x, x+h]$, and it's the j'th largest of the X_i's if and only if there are $j-1$ other X_i's that are smaller than x.

(b) Show that the probability on the right is equal to

$$n \cdot \mathbb{P}\left(X_1 \in [x, x+h]\right) \cdot \mathbb{P}\left(\text{exactly } j-1 \text{ of } X_2, X_3, \ldots, X_n \text{ are } < x\right).$$

Let A_i be the event "$X_i \in [x, x+h]$ and $j-1$ of the others are $< x$". Then

$$\begin{aligned} \mathbb{P}(A_i) &= \mathbb{P}\left((X_i \in [x, x+h]) \cap (\text{exactly } j-1 \text{ of } X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n \text{ are } < x)\right) \\ &= \mathbb{P}\left(X_i \in [x, x+h]\right) \cdot \mathbb{P}\left(\text{exactly } j-1 \text{ of } X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n \text{ are } < x\right) \end{aligned}$$

since X_i is independent of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$. Moreover, we certainly have that $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ has the same joint distribution as $(X_2, X_3, \ldots, X_n)$ because the X_i's are independent and identically distributed, which gives

$$\mathbb{P}(A_i) = \mathbb{P}\left(X_1 \in [x, x+h]\right) \cdot \mathbb{P}\left(\text{exactly } j-1 \text{ of } X_2, X_3, \ldots, X_n \text{ are } < x\right),$$

and so

$$\begin{aligned} &\mathbb{P}\left(\text{One of the } X_i\text{'s is in } [x, x+h] \text{ and exactly } j-1 \text{ of the others are } < x\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \\ &= \sum_{i=1}^{n} \mathbb{P}(A_i) && \text{Because the } A_i\text{'s are disjoint} \\ &= n \cdot \mathbb{P}\left(X_1 \in [x, x+h]\right) \cdot \mathbb{P}\left(\text{exactly } j-1 \text{ of } X_2, X_3, \ldots, X_n \text{ are } < x\right) \end{aligned}$$

(c) Think binomially and show that

$$\mathbb{P}\left(\text{exactly } j-1 \text{ of } X_2, X_3, \ldots, X_n \text{ are } < x\right) = \binom{n-1}{j-1} \cdot F_X(x)^{j-1} \cdot (1 - F_X(x))^{n-j}.$$

Let $B_i = \mathbb{1}_{X_i<x}$. Then $B_1, \ldots, B_n$ are independent Bernoulli random variables, each with probability of success $\mathbb{P}(X_i < x) = F_X(x)$. Then

$$\text{exactly } j-1 \text{ of } X_2, X_3, \ldots, X_n \text{ are } < x \iff \sum_{i=2}^{n} B_i = j-1.$$

Since $\sum_{i=2}^{n} B_i \sim \text{Bin}(n-1, F_X(x))$, the probability we want is exactly the probability that a $\text{Bin}(n-1, F(x))$ random variable equals $j-1$, which is $\binom{n-1}{j-1} \cdot F_X(x)^{j-1} \cdot (1-F_X(x))^{n-j}$.

(d) Put the pieces together, divide both sides by h, and take the limit as $h \to 0$ to get

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)! \cdot (n-j)!} \cdot f_X(x) \cdot F_X(x)^{j-1} \cdot (1-F_X(x))^{n-j}.$$

The work we've done above gives

$$\begin{aligned}\mathbb{P}\left(X_{(j)} \in [x, x+h]\right) &= n \cdot \mathbb{P}(X_1 \in [x, x+h]) \cdot \binom{n-1}{j-1} \cdot F_X(x)^{j-1} \cdot (1-F_X(x))^{n-j} \\ &= \mathbb{P}(X_1 \in [x, x+h]) \cdot \frac{n!}{(j-1)!(n-j)!} \cdot F_X(x)^{j-1} \cdot (1-F_X(x))^{n-j}.\end{aligned}$$

Writing the probabilities as differences of cdfs (i.e., $\mathbb{P}(X \in [x, x+h]) = F_X(x+h) - F_X(x)$ and so on) and dividing through by h gives

$$\frac{F_{X_{(j)}}(x+h) - F_{X_{(j)}}(x)}{h} = \frac{F_X(x+h) - F_X(x)}{h} \cdot \frac{n!}{(j-1)!(n-j)!} \cdot F_X(x)^{j-1} \cdot (1-F_X(x))^{n-j}.$$

Taking the limit on both sides as $h \to 0$ gives us exactly what we want.

*6. Let $X_1, X_2, \ldots, X_n$ be independent $\text{Unif}(0,1)$ random variables. Show $X_{(j)} \sim \text{Beta}(j, n-j+1)$, and use that fact to find $\mathbb{E}\left[X_{(j)}\right]$ and $\text{Var}\left(X_{(j)}\right)$.

Since the cdf of the $\text{Unif}(0,1)$ distribution is just $F_X(x) = x$ and the pdf is just $f_X(x) = 1$ for $x \in (0,1)$, plugging these into the result above gives

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)! \cdot (n-j)!} \cdot x^{j-1} \cdot (1-x)^{n-j} = \frac{\Gamma(n+1)}{\Gamma(j) \cdot \Gamma(n-j+1)} \cdot x^{j-1} \cdot (1-x)^{n-j}, \quad x \in (0,1),$$

which is exactly the $\text{Beta}(j, n-j+1)$ pdf. Recalling that

$$\text{B}(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1} \cdot (1-x)^{\beta-1} \, dx,$$

we get that

$$\mathbb{E}\left[X_{(j)}\right] = \frac{1}{\text{B}(j, n-j+1)} \int_0^1 x \cdot x^{j-1} \cdot (1-x)^{n-j} \, dx = \frac{\text{B}(j+1, n-j+1)}{\text{B}(j, n-j+1)} = \frac{j}{n+1}.$$

Similarly, the second moment is

$$\mathbb{E}\left[X_{(j)}^2\right] = \frac{\text{B}(j+2, n-j+1)}{\text{B}(j, n-j+1)} = \frac{j \cdot (j+1)}{(n+1) \cdot (n+2)},$$

and so

$$\text{Var}\left(X_{(j)}\right) = \mathbb{E}\left[X_{(j)}^2\right] - \mathbb{E}\left[X_{(j)}\right]^2 = \frac{j \cdot (j+1)}{(n+1) \cdot (n+2)} - \frac{j^2}{(n+1)^2} = \frac{j \cdot (n-j+1)}{(n+1)^2 \cdot (n+2)}.$$

7. What's the probability that an unbiased coin lands on heads 500 times in 1000 flips, rounded to five decimal places? You know that the exact answer is $\binom{1000}{500}0.5^{1000}$, but good luck trying to evaluate that on a calculator – you'll either end up with numerical underflow or overflow. You might think to calculate the log of that and then exponentiate it after – that will definitely help with the 0.5^{1000} part, but you'll still have to deal with $\log(1000!) - 2\log(500!)$, and you just can't evaluate either of those factorials directly. You may have heard of *Stirling's formula*, which gives an approximation of the factorial function. With a bit of hand-waving, we'll derive a simple version of it here.

   (a) Let $X_1, X_2, \ldots, X_n$ be independent $\text{Exp}(\lambda)$ random variables. Using mgfs (or anything else), show that $\sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$. This is sometimes called an *Erlang* distribution.

   The mfg of the $\text{Exp}(\lambda)$ distribution is $M_X(t) = \lambda/(\lambda - t)$ for $t < \lambda$, so the mgf of $\sum_{i=1}^n X_i$ is

$$M_{\sum_{i=1}^n X_i}(t) = \frac{\lambda^n}{(\lambda - t)^n} = \left(1 - \frac{t}{\lambda}\right)^{-n}$$

   for the same range of t, which is indeed the mgf of the $\text{Gamma}(n, \lambda)$ distribution.

   (b) Set $\lambda = 1$ and fix $x \in \mathbb{R}$. Explain why we can write

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}\left(\frac{\bar{X}_n - 1}{1/\sqrt{n}} \le x\right) \approx \phi(x)$$

   when n is large, where $\phi(x) = (\sqrt{2\pi})^{-1/2} \cdot e^{-x^2/2}$ is the standard normal pdf.

   By the central limit theorem,

$$\frac{\bar{X}_n - 1}{1/\sqrt{n}} = \frac{\bar{X}_n - \mathbb{E}[X_i]}{\sqrt{\text{Var}(X_i)/n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

   so when n is large, the term on the left is approximately $\mathcal{N}(0, 1)$-distributed. Thus

$$\mathbb{P}\left(\frac{\bar{X}_n - 1}{1/\sqrt{n}} \le x\right) \approx \int_\infty^x \phi(t)\,\mathrm{d}t,$$

   and the result follows upon differentiation (this is the hand-waving part, since we're not being very precise about what the "≈" means).

   (c) Carry out the differentiation on the left-hand side, via a u-substitution and the fundamental theorem of calculus.

   We have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}\left(\frac{\bar{X}_n - 1}{1/\sqrt{n}} \le x\right) &= \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}\left(\sum_{i=1}^n X_i \le \sqrt{n}x + n\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}x}\int_0^{\sqrt{n}x+n} \frac{1}{\Gamma(n)} t^{n-1}e^{-t}\,\mathrm{d}t && \text{Because } \textstyle\sum_{i=1}^n X_i \sim \text{Gamma}(n, 1) \\
&= \frac{\sqrt{n}}{\Gamma(n)} \cdot \frac{\mathrm{d}}{\mathrm{d}x}\int_0^x (\sqrt{n}u + n)^{n-1}e^{\sqrt{n}u - n}\,\mathrm{d}t && \text{Letting } u(t) = (t-n)/\sqrt{n} \\
&= \frac{\sqrt{n}}{\Gamma(n)} \cdot (\sqrt{n}x + n)^{n-1}e^{-(\sqrt{n}x+n)}\,\mathrm{d}t && \text{By the FTC}
\end{aligned}$$

(d) Set $x = 0$ on both sides and rearrange a bit to get

$$n! \approx \sqrt{2\pi} \cdot n^{n+\frac{1}{2}} \cdot e^{-n},$$

which is Stirling's formula.

Combining 7b and 7c gives

$$\frac{\sqrt{n}}{\Gamma(n)} \cdot (\sqrt{n}x + n)^{n-1} e^{-(\sqrt{n}x+n)}\,\mathrm{d}t \approx \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}.$$

Putting $x = 0$ gives

$$\frac{\sqrt{n}}{\Gamma(n)} \cdot n^{n-1} e^{-(\sqrt{n}x+n)}\,\mathrm{d}t \approx \frac{1}{\sqrt{2\pi}},$$

and rearranging a bit gives us Stirling's formula.

(e) Approximate (to five decimal places) the probability that an unbiased coin lands on heads 500 times in 1000 flips. I get 0.02523...

The probability is

$$\begin{aligned}
\binom{1000}{500} 0.5^{1000} &= \frac{1000!}{500!^2} \cdot 0.5^{1000} \\
&\approx \frac{\sqrt{2\pi} \cdot 1000^{1000.5} \cdot e^{-1000}}{\left(\sqrt{2\pi} \cdot 500^{500.5} \cdot e^{-500}\right)^2} \cdot 0.5^{1000} && \text{Applying Stirling's formula in the numerator and denominator} \\
&= \frac{1}{\sqrt{500\pi}} && \text{After a bit of simplification} \\
&\approx 0.02523...,
\end{aligned}$$

which is, in fact, correct to four decimal places.

8. Let $k \geq 1$ be an integer and let $\lambda > 0$. Let $X \sim \text{Gamma}\,(k, \lambda)$ (this is the Erlang distribution from Question 7a). Using mathematical induction,[1] show that the cdf X can be written as

$$\mathbb{P}\,(X \leq x) = 1 - \sum_{j=0}^{k-1} \frac{e^{-\lambda x} \cdot (\lambda x)^j}{j!}.$$

It's equivalent (but a bit less cumbersome) to show

$$\underbrace{\frac{\lambda^k}{\Gamma(k)} \int_x^\infty t^{k-1} \cdot e^{-\lambda t}\,\mathrm{d}t}_{= 1 - \mathbb{P}\,(X \leq x)} = \sum_{j=0}^{k-1} \frac{e^{-\lambda x} \cdot (\lambda x)^j}{j!},$$

so let's do that. When $k = 1$, we get

$$\frac{\lambda^1}{\Gamma(1)} \int_x^\infty t^{1-1} \cdot e^{-\lambda t}\,\mathrm{d}t = e^{-\lambda x} = \sum_{j=0}^{1-1} \frac{e^{-\lambda x} \cdot (\lambda x)^j}{j!},$$

[1] If you don't know what this is, just follow these steps: first prove the result holds for the *base case* $k = 1$. Then *assume* the result holds for any $k \in \mathbb{N}$, and show that this implies the result must also hold for $k + 1$. The *principle of mathematical induction* says that if you've done that, then you've proven the result holds for all $k \in \mathbb{N}$.

which proves the base case. Now, assume the result holds for fixed $k \in \mathbb{N}$. Then

$$\begin{aligned}
&\frac{\lambda^{k+1}}{\Gamma(k+1)} \int_x^\infty t^{(k+1)-1} \cdot e^{-\lambda t} \, \mathrm{d}t \\
&= \frac{\lambda^{k+1}}{\Gamma(k+1)} \left[-\frac{t^k \cdot e^{-\lambda t}}{\lambda} \Big|_x^\infty + \frac{k}{\lambda} \int_x^\infty t^{k-1} \cdot e^{-\lambda t} \, \mathrm{d}t \right] && \text{Using integration by parts with } u(t) = t^k \text{ and } \mathrm{d}v = e^{-\lambda t} \, \mathrm{d}t \\
&= \frac{\lambda^{k+1}}{\Gamma(k+1)} \left[-\frac{t^k \cdot e^{-\lambda t}}{\lambda} \Big|_x^\infty + \frac{k}{\lambda} \cdot \frac{\Gamma(k)}{\lambda^k} \sum_{j=0}^{k-1} \frac{e^{-\lambda x} \cdot (\lambda x)^j}{j!} \right] && \text{By the induction hypothesis} \\
&= \frac{(\lambda x)^k e^{-\lambda x}}{\Gamma(k+1)} + \sum_{j=0}^{k-1} \frac{e^{-\lambda x} \cdot (\lambda x)^j}{j!} \\
&= \sum_{j=0}^{k} \frac{e^{-\lambda x} \cdot (\lambda x)^j}{j!},
\end{aligned}$$

so the result also holds for $k+1$. By the principle of mathematical induction, the result is true for all $k \in \mathbb{N}$.

9. Let $U_1, U_2, \ldots, U_n, V$ be independent $\mathrm{Unif}(0,1)$ random variables, where $n \geq 2$. Find the pdf of $Z = \left(\prod_{i=1}^n U_i\right)^V$.

   Hint: start by finding the distribution of $-\log(Z)$. This *might* be the toughest (or at least the longest) question of the batch. For an easier version, try to solve it for the $n = 2$ case.

To find $\mathbb{P}(Z \leq x)$ for $x \in (0,1)$, first let $y = -\log(x)$. Then

$$\mathbb{P}(Z \leq x) = \mathbb{P}\left(V \cdot \sum_{i=1}^n \log(U_i) \leq -y \right) = \mathbb{P}\left(V \cdot \sum_{i=1}^n (-\log(U_i)) \geq y \right).$$

Let's find the distribution of $\sum_{i=1}^n (-\log(U_i))$. We see that $-\log(U_i) \sim \mathrm{Exp}(1)$ because $\mathbb{P}(-\log(U_i) \leq u) = \mathbb{P}(U_i \geq e^{-u}) = 1 - e^{-u}$, so $\sum_{i=1}^n (-\log(U_i))$ is a sum of n independent $\mathrm{Exp}(1)$ random variables, which gives $\sum_{i=1}^n (-\log(U_i)) \sim \mathrm{Gamma}(n,1)$ by Question 7a. So we can write $\mathbb{P}(Z \leq x) = 1 - \mathbb{P}(V \cdot G < y)$, where $V \sim \mathrm{Unif}(0,1)$ and $G \sim \mathrm{Gamma}(n,1)$ are independent. Now, using the law of total probability,

$$\begin{aligned}
&\mathbb{P}(V \cdot G < y) \\
&= \int_0^\infty \mathbb{P}(V \cdot G < y \mid G = g) \cdot f_G(g) \, \mathrm{d}g \\
&= \int_0^\infty \mathbb{P}\left(V < \frac{y}{g} \right) \cdot \frac{g^{n-1} \cdot e^{-g}}{\Gamma(n)} \, \mathrm{d}g \\
&= \int_0^y \mathbb{P}\left(V < \frac{y}{g} \right) \cdot \frac{g^{n-1} \cdot e^{-g}}{\Gamma(n)} \, \mathrm{d}g + \int_y^\infty \mathbb{P}\left(V < \frac{y}{g} \right) \cdot \frac{g^{n-1} \cdot e^{-g}}{\Gamma(n)} \, \mathrm{d}g \\
&= \int_0^y \frac{g^{n-1} \cdot e^{-g}}{\Gamma(n)} \, \mathrm{d}g + \frac{y}{\Gamma(n)} \int_y^\infty g^{n-2} \cdot e^{-g} \, \mathrm{d}g && \text{Since } V \sim \mathrm{Unif}(0,1), \text{ so } \mathbb{P}(V < a) = a \cdot \mathbb{1}_{0 \leq a < 1} + 1 \cdot \mathbb{1}_{a \geq 1} \\
&= \int_0^y \frac{g^{n-1} \cdot e^{-g}}{\Gamma(n)} \, \mathrm{d}g + y \cdot \frac{\Gamma(n-1)}{\Gamma(n)} \int_y^\infty \frac{g^{n-2} \cdot e^{-g}}{\Gamma(n-1)} \, \mathrm{d}g \\
&= \mathbb{P}(G \leq y) + \frac{y}{n-1} \cdot (1 - \mathbb{P}(H \leq y)) && \text{Where } H \sim \mathrm{Gamma}(n-1,1)
\end{aligned}$$

Therefore,

$$\mathbb{P}(Z \leq x) = 1 - \mathbb{P}(G \leq y) - \frac{y}{n-1} \cdot (1 - \mathbb{P}(H \leq y)).$$

To get the pdf, we can use the chain rule and the result from Question 8:

$$\begin{aligned}
f_Z(x) &= \frac{\mathrm{d}y}{\mathrm{d}x} \cdot \frac{\mathrm{d}}{\mathrm{d}y}\left(1 - \mathbb{P}\left(G \le y\right) - \frac{y}{n-1} \cdot \left(1 - \mathbb{P}\left(H \le y\right)\right)\right) \\
&= \frac{\mathrm{d}y}{\mathrm{d}x} \cdot \left(-f_G(y) - \left[\frac{1 - \mathbb{P}\left(H \le y\right)}{n-1} - \frac{y}{n-1} \cdot f_H(y)\right]\right) \\
&= \frac{1}{e^{-y}} \cdot \left(f_G(y) + \frac{1 - \mathbb{P}\left(H \le y\right)}{n-1} - \frac{y}{n-1} \cdot f_H(y)\right) \\
&= \frac{1}{e^{-y}} \cdot \left(\frac{y^{n-1} \cdot e^{-y}}{\Gamma(n)} + \frac{1}{n-1} \cdot \sum_{j=0}^{n-2} \frac{e^{-y} \cdot y^j}{j!} - \frac{y}{n-1} \cdot \frac{y^{n-2} \cdot e^{-y}}{\Gamma(n-1)}\right) \\
&= \frac{y^{n-1}}{\Gamma(n)} + \frac{1}{n-1} \cdot \sum_{j=0}^{n-2} \frac{y^j}{j!} - \frac{1}{n-1} \cdot \frac{y^{n-1}}{\Gamma(n-1)} \\
&= \frac{1}{n-1} \sum_{j=0}^{n-2} \frac{y^j}{j!} \\
&= \frac{1}{n-1} \sum_{j=0}^{n-2} \frac{(-1)^j \cdot \log\left(x\right)^j}{j!},
\end{aligned}$$

which is good enough.[2]

10. Let $U_1, U_2, \ldots$ be independent $\mathrm{Unif}\left(0,1\right)$ random variables. Let M be a random variable independent of the U_i's, with distribution

$$\mathbb{P}\left(M = m\right) = \frac{c}{m!}, \quad m = 1, 2, 3, \ldots$$

for some $c \in \mathbb{R}$. Find the value of c, and then find the pdf of $X = \min\{U_1, U_2, \ldots, U_M\}$. That's the minimum of a random number of U_i's, so you'll have to do some kind of conditioning.

First of all, we need

$$1 = \sum_{m=1}^{\infty} \frac{c}{m!} = c(e-1),$$

which gives $c = 1/(e-1)$. Now, by the law of total probability, for $x \in (0,1)$ the cdf of X is

$$\begin{aligned}
\mathbb{P}\left(X \le x\right) &= \sum_{m=1}^{\infty} \mathbb{P}\left(\min\{U_1, U_2, \ldots, U_M\} \le x \mid M = m\right) \cdot \mathbb{P}\left(M = m\right) \\
&= \sum_{m=1}^{\infty} \mathbb{P}\left(\min\{U_1, U_2, \ldots, U_m\} \le x\right) \cdot \mathbb{P}\left(M = m\right) \\
&= \frac{1}{e-1} \sum_{m=1}^{\infty} \frac{1 - (1-x)^m}{m!} \\
&= \frac{1}{e-1} \sum_{m=0}^{\infty} \frac{1 - (1-x)^m}{m!} \\
&= \frac{e - e^{1-x}}{e-1},
\end{aligned}$$

[2] Apparently the $n = 2$ case, which you can see gives $(U_1 \cdot U_2)^V \sim \mathrm{Unif}\left(0,1\right)$, has been known to be given as an interview question by some hedge funds. So now you're ready to be a quant!

and the pdf of X is the derivative of that, which is $e^{1-x}/(e-1)$.

11. Suppose you repeatedly draw independent $\text{Unif}(0,1)$ random variables and add them together. What's the expected number of draws you need for the sum to exceed 1? Let's answer that.

    (a) Let $U_1, U_2, \ldots, U_n$ be independent $\text{Unif}(0,1)$ random variables, and let $S_n = \sum_{i=1}^n U_i$. Using mathematical induction, prove that $\mathbb{P}(S_k \le t) = t^k/k!$ for $t \in (0,1)$.

    When $k = 1$, we have that $\mathbb{P}(S_1 \le t) = \mathbb{P}(U_1 \le t) = t = t^1/1!$, which proves the base case. Now, assume the result holds for fixed $k \in \mathbb{N}$. Then, by the law of total probability,

$$\begin{aligned}
\mathbb{P}(S_{k+1} \le t) &= \mathbb{P}(S_k \le t - U_{k+1}) \\
&= \int_0^1 \mathbb{P}(S_k \le t - u \mid U_{k+1} = u) \cdot \mathbb{1}_{t-u>0}\, du \\
&= \int_0^t \mathbb{P}(S_k \le t - u)\, du && \text{Since } S_k \text{ is independent of } U_{k+1} \\
&= \int_0^t \frac{(t-u)^k}{k!}\, du && \text{By the induction hypothesis} \\
&= \frac{t^{k+1}}{(k+1)!},
\end{aligned}$$

    so the result also holds for $k+1$. By the principle of mathematical induction, the result is true for all $k \in \mathbb{N}$.

    (b) Let $N = \min\{k : S_k > 1\}$. Argue that $\mathbb{P}(N = n) = \mathbb{P}(S_{n-1} \le 1) - \mathbb{P}(S_n \le 1)$.

    $N = n$ happens if and only if both $S_n > 1$ and $S_{n-1} \le 1$ happen simultaneously. In other words, $\{N = n\} = \{S_n > 1\} \cap \{S_{n-1} \le 1\} = \{S_n \le 1\}^c \cap \{S_{n-1} \le 1\}$. The result now follows from the basic property $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$ whenever $A \subseteq B$.

    (c) Use that to evaluate $\mathbb{E}[N]$. Think about where your summation starts!

    Because $S_1 = U_1$ definitely can't exceed 1, we have $\mathbb{P}(N = 1) = 0$. Therefore,

$$\begin{aligned}
\mathbb{E}[N] &= \sum_{n \ge 2} n \cdot \mathbb{P}(N = n) \\
&= \sum_{n \ge 2} n \cdot \left(\frac{1}{(n-1)!} - \frac{1}{n!} \right) \\
&= \sum_{n \ge 2} \frac{1}{(n-2)!} \\
&= \sum_{j \ge 0} \frac{1}{j!} && \text{Substituting } j = n - 2 \\
&= e.
\end{aligned}$$

    So we "expect" to draw exactly e independent standard uniform random variables before their sum exceeds 1 (!).

12. If $\mathbf{X} = (X_1, X_2, X_3, X_4)$ is jointly distributed according to

$$f_{\mathbf{X}}(x_1, x_2, x_3, x_4) = \frac{3}{4}(x_1^2 + x_2^2 + x_3^2 + x_4^2), \quad 0 < x_i < 1, \qquad i = 1, 2, 3, 4,$$

find $\mathbb{P}\left(X_1 < \sqrt{X_2} < X_3 < \sqrt{X_4}\right)$ and $\mathbb{E}\left[\sqrt{X_1} \cdot X_3\right]$.

The probability is just

$$\mathbb{P}\left(X_1 < \sqrt{X_2} < X_3 < \sqrt{X_4}\right) = \int_0^1 \int_0^{\sqrt{x_4}} \int_0^{x_3} \int_0^{\sqrt{x_2}} f_{\mathbf{X}}(x_1, x_2, x_3, x_4)\, \mathrm{d}x_1\, \mathrm{d}x_2\, \mathrm{d}x_3\, \mathrm{d}x_4 = \frac{81{,}392}{765{,}765}$$

while the expectation is

$$\mathbb{E}\left[\sqrt{X_1} \cdot X_3\right] = \int_0^1 \int_0^1 \sqrt{x_1} \cdot x_3 \left(\int_0^1 \int_0^1 f_{\mathbf{X}}(x_1, x_2, x_3, x_4)\, \mathrm{d}x_2\, \mathrm{d}x_4 \right) \mathrm{d}x_1\, \mathrm{d}x_3 = \frac{67}{168}.$$

13. Let B and C be independent $\text{Unif}\,(0, 1)$ random variables. Find the probability that the random quadratic $x^2 + Bx + C$ has a real root. For a harder version, let $A \sim \text{Unif}\,(0, 1)$ be independent of B and C and find the probability that $Ax^2 + Bx + C$ has a real root.

From the quadratic formula, the general quadratic $x^2 + bx + c$ has a real root if and only if the discriminant $\sqrt{b^2 - 4c}$ is real, which itself happens if and only if $b^2 - 4c \geq 0$. So we want

$$\begin{aligned}
\mathbb{P}\left(B^2 - 4C \geq 0\right) &= \int_0^1 \mathbb{P}\left(B^2 - 4C \geq 0 \mid B = b\right) \mathrm{d}b \\
&= \int_0^1 \mathbb{P}\left(C \leq \frac{b^2}{4}\right) \mathrm{d}b \\
&= \int_0^1 \frac{b^2}{4}\, \mathrm{d}b && \text{Since } C \sim \text{Unif}\,(0, 1) \\
&= \frac{1}{12}.
\end{aligned}$$

For the harder version, let $G \sim \text{Gamma}\,(2, 1)$. This time we want

$$\int_0^1 \mathbb{P}\left(AC \leq \frac{b^2}{4}\right) \mathrm{d}b = \int_0^1 \mathbb{P}\left(-\log(A) - \log(C) \geq -\log\left(\frac{b^2}{4}\right)\right) \mathrm{d}b = \int_0^1 \mathbb{P}\left(G \geq -\log\left(\frac{b^2}{4}\right)\right) \mathrm{d}b,$$

and using Question 8, that's

$$\begin{aligned}
&\int_0^1 \left(e^{\log(b^2/4)} - e^{\log(b^2/4)} \cdot \log\left(\frac{b^2}{4}\right)\right) \mathrm{d}b \\
&= \int_0^1 \frac{b^2}{4}\, \mathrm{d}b - \int_0^1 \frac{b^2}{2} \cdot \log\left(\frac{b}{2}\right) \mathrm{d}b \\
&= \int_0^1 \frac{b^2}{4}\, \mathrm{d}b - \left[\frac{b^3}{6} \cdot \log\left(\frac{b}{2}\right)\Big|_0^1 - \int_0^1 \frac{b^2}{6}\, \mathrm{d}b\right] && \text{Using integration by parts on the second integral with } u(b) = \log\left(\tfrac{b}{2}\right) \text{ and } \mathrm{d}v = \tfrac{b^2}{2}\, \mathrm{d}b \\
&= \frac{1}{12} - \frac{1}{6} \cdot \log\left(\frac{1}{2}\right) + \frac{1}{18} && \text{Using L'Hôpital's rule to get } \lim_{b \to 0} \tfrac{b^3}{6} \cdot \log\left(\tfrac{b}{2}\right) = 0 \\
&= \frac{5}{36} + \frac{\log(2)}{6}.
\end{aligned}$$

⋆14. Let Y be a random variable whose first two moments exist. Hypothesize which $x \in \mathbb{R}$ minimizes $\mathbb{E}\left[(Y-x)^2\right]$, and then prove it.

The minimizing value of x is $x = \mathbb{E}[Y]$, so that $\inf_{x\in\mathbb{R}} \mathbb{E}\left[(Y-x)^2\right] = \mathrm{Var}(Y)$. To prove it, expand the square to get

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left[(Y-x)^2\right] = \frac{\mathrm{d}}{\mathrm{d}x}\left(\mathbb{E}\left[Y^2\right] - 2x\mathbb{E}[Y] + x^2\right) = -2\mathbb{E}[Y] + 2x.$$

Setting the right-hand side equal to 0 shows that $x = \mathbb{E}[Y]$ is a global optimizer. Differentiating the display above one more time shows that we always have $\frac{\mathrm{d}^2}{\mathrm{d}x^2}\mathbb{E}\left[(Y-x)^2\right] = 2 > 0$, which confirms that $x = \mathbb{E}[Y]$ is indeed the global minimum.

15. Let $X \sim \mathrm{Poisson}(\lambda)$ and $Y \sim \mathrm{Poisson}(\nu)$ be independent. Find the conditional distribution of $X \mid (X+Y=n)$.

First of all, it's easy to show using mgfs that $X+Y \sim \mathrm{Poisson}(\lambda+\nu)$. Then for $m \in \mathbb{N}$ with $m \le n$, we have

$$\begin{aligned}
\mathbb{P}(X=m \mid X+Y=n) &= \frac{\mathbb{P}(X=m \wedge X+Y=n)}{\mathbb{P}(X+Y=n)} \\
&= \frac{\mathbb{P}(X=m \wedge Y=n-m)}{\mathbb{P}(X+Y=n)} \\
&= \frac{\mathbb{P}(X=m)\cdot\mathbb{P}(Y=n-m)}{\mathbb{P}(X+Y=n)} && \text{Since } X \text{ and } Y \text{ are independent} \\
&= \frac{\lambda^m e^{-\lambda}}{m!}\cdot\frac{\nu^{n-m}e^{-\nu}}{(n-m)!}\Bigg/\frac{(\lambda+\nu)^n e^{-(\lambda+\nu)}}{n!} \\
&= \binom{n}{m}\left(\frac{\lambda}{\lambda+\nu}\right)^m\left(1-\frac{\lambda}{\lambda+\nu}\right)^{n-m}.
\end{aligned}$$

That is, $X \mid (X+Y=n) \sim \mathrm{Bin}(n, \lambda/(\lambda+\nu))$.

16. Let $X \sim \mathrm{Gamma}(\lambda,1)$ and $Y \sim \mathrm{Gamma}(\nu,1)$ be independent. Name the distributions of $G = X+Y$ and $B = X/(X+Y)$, and show they're independent. Don't try to start by finding the marginals – instead, go straight for the joint distribution of (G,B) and see what pops out.

Let $g = g(x,y) = x+y$ and $h = h(x,y) = x/(x+y)$. Then the function $(x,y) \mapsto (g(x,y), h(x,y))$ is a smooth bijection between $(0,\infty)^2$ and $(0,\infty)\times(0,1)$ with inverse $(g,h) \mapsto (x(g,h), y(g,h)) = (gh, g(1-h))$. The determinant of the Jacobian of the inverse transformation is

$$\det\left(\frac{\mathrm{d}(x,y)}{\mathrm{d}(g,h)}\right) = \left|\begin{bmatrix}\frac{\partial x}{\partial g} & \frac{\partial x}{\partial h} \\ \frac{\partial y}{\partial g} & \frac{\partial y}{\partial h}\end{bmatrix}\right| = \left|\begin{bmatrix} h & g \\ 1-h & -g\end{bmatrix}\right| = g,$$

so the joint pdf of (G,H) is

$$\begin{aligned}
f_{(G,H)}(g,h) &= f_{(X,Y)}(gh, g(1-h))\cdot\left|\det\left(\frac{\mathrm{d}(x,y)}{\mathrm{d}(g,h)}\right)\right| \\
&= f_X(gh)\cdot f_Y(g(1-h))\cdot\left|\det\left(\frac{\mathrm{d}(x,y)}{\mathrm{d}(g,h)}\right)\right| && \text{Since } X \text{ and } Y \text{ are independent}
\end{aligned}$$

$$= \frac{1}{\Gamma(\lambda)}(gh)^{\lambda-1}e^{-gh} \cdot \frac{1}{\Gamma(\nu)}(g(1-h))^{\nu-1}e^{-g(1-h)} \cdot g$$
$$= \underbrace{\frac{1}{\Gamma(\lambda+\nu)}g^{\lambda+\nu-1}e^{-g}}_{\text{Gamma}\,(\lambda+\nu,1)\text{ pdf}} \cdot \underbrace{\frac{\Gamma(\lambda+\nu)}{\Gamma(\lambda)\cdot\Gamma(\nu)}h^{\lambda-1}(1-h)^{\nu-1}}_{\text{Beta}\,(\lambda,\nu)\text{ pdf}}.$$

So $G \sim \text{Gamma}\,(\lambda+\nu,1)$ and $B \sim \text{Beta}\,(\lambda,\rho)$ and moreover, the factorization above shows that G and B are independent.

17. Let X and Y be independent $\mathcal{N}(0,1)$ random variables.

(a) Let $R = \sqrt{X^2+Y^2}$ and $\Theta = \arctan\left(\frac{Y}{X}\right)$, where the range of arctan is taken as $[0,2\pi]$. Name the distributions of R^2 and Θ, and show they're independent. Again, go straight for their joint distribution. If your trig is rusty, remember that $\tan(x) = \sin(x)/\cos(x)$ and $\sin^2(x)+\cos^2(x) = 1$.

Let $r = r(x,y) = \sqrt{x^2+y^2}$ and $\theta = \theta(x,y) = \arctan(y/x)$. Then the function $(x,y) \mapsto (r(x,y),\theta(x,y))$ is a smooth bijection between $\mathbb{R}^2$ and $[0,\infty)\times[0,2\pi]$ with inverse $(r,\theta) \mapsto (x(r,\theta),y(r,\theta)) = (r\cdot\cos(\theta), r\cdot\sin(\theta))$. The determinant of the Jacobian of the inverse transformation is

$$\det\left(\frac{\mathrm{d}(x,y)}{\mathrm{d}(r,\theta)}\right) = \left|\begin{bmatrix}\frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta}\\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta}\end{bmatrix}\right| = \left|\begin{bmatrix}\cos(\theta) & -r\cdot\sin(\theta)\\ \sin(\theta) & r\cdot\cos(\theta)\end{bmatrix}\right| = r,$$

so the joint pdf of (R,Θ) is

$$\begin{aligned} f_{(R,\Theta)}(r,\theta) &= f_{(X,Y)}(r\cdot\cos(\theta), r\cdot\sin(\theta)) \cdot \left|\det\left(\frac{\mathrm{d}(x,y)}{\mathrm{d}(r,\theta)}\right)\right| \\ &= f_X(r\cdot\cos(\theta)) \cdot f_Y(r\cdot\sin(\theta)) \cdot \left|\det\left(\frac{\mathrm{d}(x,y)}{\mathrm{d}(r,\theta)}\right)\right| \quad \text{Since } X \text{ and } Y \text{ are independent} \\ &= \frac{1}{\sqrt{2\pi}}e^{-r^2\cdot\cos(\theta)^2/2} \cdot \frac{1}{\sqrt{2\pi}}e^{-r^2\cdot\sin(\theta)^2/2} \cdot r \\ &= \frac{1}{2\pi}\cdot re^{-r^2/2}. \end{aligned}$$

Since the marginal pdf of Θ is $\int_0^\infty f_{(R,\Theta)}(r,\theta)\,\mathrm{d}r = 1/2\pi$, it follows that $\Theta \sim \text{Unif}\,(0,2\pi)$, and also that R and Θ are independent. Moreover, by a change of variables, the marginal pdf of R^2 is $f_{R^2}(x) = e^{-x/2}/2$ which gives us $R^2 \sim \text{Exp}\,(1/2)$.

(b) Use your work to show that if U_1 and U_2 are independent $\text{Unif}\,(0,1)$ random variables, then $X \overset{d}{=} \sqrt{-2\log(U_1)}\cdot\cos(2\pi U_2)$ and $Y \overset{d}{=} \sqrt{-2\log(U_1)}\cdot\sin(2\pi U_2)$. This is called the *Box-Muller transform*.

If $U_1, U_2 \sim \text{Unif}\,(0,1)$, then $\mathbb{P}\left(\sqrt{-2\log(U_1)} \le r\right) = \mathbb{P}\left(U_1 \ge e^{-r^2/2}\right) = 1 - e^{-r^2/2}$, and taking derivatives shows that $\sqrt{-2\log(U_1)}$ has pdf $re^{-r^2/2}$, so that $\sqrt{-2\log(U_1)} \overset{d}{=} R$. Also, we certainly have that $2\pi\cdot U_2 \sim \text{Unif}\,(0,2\pi)$; that is $2\pi U_2 \overset{d}{=} \Theta$. Putting the pieces together and using Question 17a gives us

$$X \overset{d}{=} R\cdot\cos(\Theta) \overset{d}{=} \sqrt{-2\log(U_1)}\cdot\cos(2\pi U_2)$$

and

$$Y \stackrel{d}{=} R \cdot \cos(\Theta) \stackrel{d}{=} \sqrt{-2\log(U_1)} \cdot \sin(2\pi U_2).$$

(c) If I give you only a pocket calculator and two independent draws from the $\text{Unif}(0,1)$ distribution, explain how you can give me back independent draws from the $\mathcal{N}(\mu_1, \sigma_1^2)$ distribution and the $\mathcal{N}(\mu_2, \sigma_2^2)$ distribution.

If I give you u_1 and u_2 (assumed to be independent $\text{Unif}(0,1)$ realizations), you give me back $\sigma_1 \cdot \sqrt{-2\log(u_1)} \cdot \cos(2\pi u_2) + \mu_1$ and $\sigma_2 \cdot \sqrt{-2\log(u_1)} \cdot \sin(2\pi u_2) + \mu_2$.

18. Let X_1, X_2 and X_3 be uncorrelated random variables, all with expectation μ and variance σ^2. Find expressions for $\text{Cov}(X_1 + X_2, X_2 + X_3)$ and $\text{Cov}(X_1 + X_2, X_1 - X_2)$ in terms of μ and σ^2.

Using independence and the identity $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, we get

$$\begin{aligned}
&\text{Cov}(X_1 + X_2, X_2 + X_3) \\
&= \mathbb{E}[(X_1 + X_2) \cdot (X_2 + X_3)] - \mathbb{E}[X_1 + X_2] \cdot \mathbb{E}[X_2 + X_3] \\
&= \mathbb{E}[X_1 \cdot X_2 + X_1 \cdot X_3 + X_2^2 + X_2 \cdot X_3] - \mathbb{E}[X_1 + X_2] \cdot \mathbb{E}[X_2 + X_3] \\
&= \mathbb{E}[X_1] \cdot \mathbb{E}[X_2] + \mathbb{E}[X_1] \cdot \mathbb{E}[X_3] + \mathbb{E}[X_2^2] + \mathbb{E}[X_2] \cdot \mathbb{E}[X_3] - \mathbb{E}[X_1 + X_2] \cdot \mathbb{E}[X_2 + X_3] \\
&= \sigma^2.
\end{aligned}$$

Also,

$$\begin{aligned}
\text{Cov}(X_1 + X_2, X_1 - X_2) &= \mathbb{E}[(X_1 + X_2) \cdot (X_1 - X_2)] - \mathbb{E}[X_1 + X_2] \cdot \overbrace{\mathbb{E}[X_1 - X_2]}^{0} \\
&= \mathbb{E}[X_1^2 - X_2^2] \\
&= \text{Var}(X_1) + \mathbb{E}[X_1]^2 - \left(\text{Var}(X_2) + \mathbb{E}[X_2]^2\right) \\
&= 0
\end{aligned}$$

⋆19. Let $X_1, X_2, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Define the *sample mean* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and the *sample variance* $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Prove that $\mathbb{E}[\bar{X}_n] = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$ and also $\mathbb{E}[S_n^2] = \sigma^2$.
Hint: for the last one, you can make life easier by writing $X_i - \bar{X}_n = (X_i - \mu) - (\bar{X}_n - \mu)$.

To begin with, by linearity we have

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^n \mu = \mu.$$

Moreover, because the X_i are independent, we also have

$$\text{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}.$$

To show that $\mathbb{E}[S_n^2] = \sigma^2$, it's obviously enough to show that $\sum_{i=1}^n \mathbb{E}[(X_i - \bar{X}_n)^2] = (n-1)\cdot\sigma^2$. Taking the hint, we see that

$$\sum_{i=1}^n \mathbb{E}[(X_i - \bar{X}_n)^2] = \sum_{i=1}^n \mathbb{E}\left[\left((X_i - \mu) - (\bar{X}_n - \mu)\right)^2\right]$$

$$= \sum_{i=1}^{n} \left(\mathbb{E}\left[(X_i - \mu)^2\right] - 2\mathbb{E}\left[(X_i - \mu) \cdot (\bar{X}_n - \mu)\right] + \mathbb{E}\left[(\bar{X}_n - \mu)^2\right] \right)$$
$$= (n-1) \cdot \sigma^2$$

because $\mathbb{E}\left[(X_i - \mu)^2\right] = \mathrm{Var}\left(X_i\right)$ and $\mathbb{E}\left[(\bar{X}_n - \mu)^2\right] = \mathrm{Var}\left(\bar{X}_n\right)$ and

$$\sum_{i=1}^{n} \mathbb{E}\left[(X_i - \mu) \cdot (\bar{X}_n - \mu)\right] = \mathbb{E}\left[(\bar{X}_n - \mu) \cdot \sum_{i=1}^{n} (X_i - \mu)\right] = n \cdot \mathbb{E}\left[(\bar{X}_n - \mu)^2\right] = n \cdot \mathrm{Var}\left(\bar{X}_n\right).$$

20. In the same setting as above, show that the sample variance satisfies

$$S_n^2 = \frac{n-2}{n-1} S_{n-1}^2 + \frac{1}{n}(\bar{X}_{n-1} - X_n)^2.$$

Why might this identity be useful?

Hint: add and subtract $\bar{X}_{n-1}$ inside the summands being squared in S_n^2.

Following the hint, we have

$$\begin{aligned}
(n-1)S_n^2 &= \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \\
&= \sum_{i=1}^{n} \left((X_i - \bar{X}_{n-1}) + (\bar{X}_{n-1} - \bar{X}_n) \right)^2 \\
&= \sum_{i=1}^{n} (X_i - \bar{X}_{n-1})^2 + 2\sum_{i=1}^{n} (X_i - \bar{X}_{n-1}) \cdot (\bar{X}_{n-1} - \bar{X}_n) + \sum_{i=1}^{n} (\bar{X}_{n-1} - \bar{X}_n)^2 \\
&= \sum_{i=1}^{n} (X_i - \bar{X}_{n-1})^2 + 2n(\bar{X}_n - \bar{X}_{n-1}) \cdot (\bar{X}_{n-1} - \bar{X}_n) + n(\bar{X}_{n-1} - \bar{X}_n)^2 \\
&= \sum_{i=1}^{n} (X_i - \bar{X}_{n-1})^2 - n(\bar{X}_{n-1} - \bar{X}_n)^2
\end{aligned}$$

The first term is

$$(n-2)S_{n-1}^2 + (\bar{X}_{n-1} - X_n)^2,$$

while the second is

$$-n\left(\frac{1}{n-1}\sum_{i=1}^{n-1} X_i - \frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = -n\left(\frac{1}{n \cdot (n-1)}\sum_{i=1}^{n-1} X_i - \frac{1}{n} X_n\right)^2 = -\frac{1}{n}(\bar{X}_{n-1} - X_n)^2.$$

Adding those gives

$$(n-1)S_n^2 = (n-2)S_{n-1}^2 + \frac{n-1}{n}(\bar{X}_{n-1} - X_n)^2,$$

and dividing through by $n-1$ gives us what we want.

From a computational perspective, this is very useful because if we've already calculated a sample mean and sample variance based on $n-1$ data points $X_1, \ldots, X_{n-1}$, and we're given a new point X_n to add into the mix, we don't have to go back and recalculate the updated sample variance from scratch, which would normally require $O(n)$ operations; instead, the recursive identity we just derived does it in a *constant* number of operations, which saves a huge amount of time if n is large.

21. Let $\mathbf{A}$ be an $n \times n$ matrix whose entries are independent $\mathcal{N}(0,1)$ random variables. Let $\mathbf{B} = (\mathbf{A}+\mathbf{A}^\top)/2$, which you might notice is symmetric. What's the joint pdf of the $n(n+1)/2$ entries in the upper triangle of B? This has matrices in it, but it doesn't need any linear algebra; if you remember what the transpose of a matrix is, you can do this! If you're looking for a name for your pdf, you can call it $f_{B_{11},B_{12},\ldots,B_{nn}}(b_{11}, b_{12}, \ldots, b_{nn})$.

Let B_{ij} be the (i,j)'th entry of $\mathbf{B}$. If $i = j$, then $B_{ii} = (A_{ii} + A_{ii})/2 = A_{ii} \sim \mathcal{N}(0,1)$. On the other hand, if $i \neq j$, then $B_{ij} = (A_{ij} + A_{ji})/2 \sim \mathcal{N}(0, 1/2)$. If $1 \leq i \leq j \leq n$, then all of these B_{ij}'s are independent, so their joint pdf is just

$$\begin{aligned} f_{B_{11},B_{12},\ldots,B_{nn}}(b_{11}, b_{12}, \ldots, b_{nn}) &= \overbrace{\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-b_{ii}^2/2}\right)}^{\text{diagonal entries}} \cdot \overbrace{\left(\prod_{i=1}^{n}\prod_{j\neq i} \frac{1}{\sqrt{\pi}} e^{-b_{ij}^2}\right)}^{\text{off-diagonal entries}} \\ &= \frac{1}{2^{n/2} \cdot \pi^{n(n-1)/4}} \cdot \exp\left(-\sum_{i=1}^{n}\left[\frac{b_{ii}^2}{2} + \sum_{j\neq i} b_{ij}^2\right]\right). \end{aligned}$$

22. Fix some $n \in \mathbb{N}$ with $n > 1$. Prove that if I give you some fixed $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, you can give me $x_1, x_2, \ldots, x_n \in \mathbb{R}$ such that

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = \mu$$

and

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sigma^2.$$

What — if any — are some statistical implications of this?

Hint: start with $n = 2$, and you'll get an explicit form for x_1 and x_2. Use those to take a guess at the case for general $2n$, and prove that it gives you what you want. For odd n, add an appropriate x_{2n+1} to the $2n$ case.

For the $n = 2$ case, it's possible to expand everything out and solve the simultaneous equations, but it's much easier to go by intuition. The endpoints of any interval centered at μ will have μ as their average, so we can take $x_1 = \mu - q$ and $x_2 = \mu + q$ for some $q > 0$ (which gives $\bar{x} = \mu$ like we want), and plugging this into the second formula gives $q = \sqrt{\sigma^2/2}$.

For the general case where n is even, follow the same strategy: take $x_{2j} = \mu - q$ and $x_{2j-1} = \mu + q$ for each $j = 1, \ldots, n/2$ so that $\bar{x} = \mu$ for any q. With these choices, we want

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\pm q)^2 = \frac{n}{n-1}q^2,$$

which means that we can take $q = \sqrt{(1 - 1/n)\sigma^2}$.

Finally, if n is odd – say $n = 2m + 1$ – one thing we can do is let $x_1, \ldots, x_{2m}$ be as above (for some q to be specified) and let $x_{2m+1} = \mu$. Then we still have $\bar{x} = \mu$, so that $(x_{2m+1} - \bar{x})^2 = 0$. Plugging this in shows $q = \sqrt{\sigma^2}$ does the trick.

One statistical implication here is that given any $n > 1$, someone can easily cook up a dataset of size n that has any prescribed sample mean and sample variance. Of course, if the data is supposed to be collected from "real-life" then this dataset will not be very convincing (since

it will contain at most three unique values) but someone only interested in the two summary statistics above will not necessarily notice this.

23. In STA257, you learned *Chebyshev's inequality*, a corollary of which says that if $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then $\mathbb{P}(|X - \mu| \geq \lambda) \leq \sigma^2/\lambda^2$ for any $\lambda > 0$. This is the most basic example of a *concentration inequality*, so named because it essentially says that random variables with finite moments tend to "concentrate" around their means — in this case, the probability that X is at a distance at least x away from μ decays like $1/x^2$. It turns out that Chebyshev's inequality is often rather weak, and for sums of nice independent random variables, we can obtain much stronger concentration.

    (a) First show that Chebyshev's inequality is tight (i.e., equality holds for some random variable X and some $\lambda > 0$). The easiest example is discrete — try and construct X so that it gives you what you need.

    For example, take X such that $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Then $\mathbb{E}[X] = 0$ and $\mathrm{Var}(X) = 1$ so that

    $$1 = \mathbb{P}(|X| \geq 1) = \mathbb{P}(|X - \mathbb{E}[X]| \geq 1) \leq \frac{\mathrm{Var}(X)}{1^2} = 1,$$

    so Chebyshev's inequality is tight here with $\lambda = 1$. You can do the same kind of trick (with a slight modification) for any fixed $\lambda > 0$.

    (b) Let $X_i \sim \text{Bernoulli}(p_i)$ be independent for $i = 1, \ldots, n$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \sum_{i=1}^n p_i$.

        i. Let $M_X(t)$ be the mgf of X. Use the fact that $1 + x \leq e^x$ to show that $M_X(t) \leq e^{\mu(e^t - 1)}$.

        Since the mgf of the $\text{Bernoulli}(p_i)$ distribution is $M_{X_i}(t) = 1 + p_i(e^t - 1)$, using independence of the X_i's and the provided bound, we have

        $$M_X(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n \left(1 + p_i(e^t - 1)\right) \leq \prod_{i=1}^n e^{p_i(e^t - 1)} = e^{\mu(e^t - 1)}.$$

        ii. Use Markov's inequality and the inequality above to show that for any $\delta > 0$ and any $t \in \mathbb{R} \setminus \{0\}$,

        $$\mathbb{P}(X \geq \mu(1 + \delta)) \leq \left(\frac{e^{(e^t - 1)}}{e^{t(1+\delta)}}\right)^{\mu}.$$

        Using Markov's inequality and the bound above, we get

        $$\mathbb{P}(X \geq \mu(1 + \delta)) = \mathbb{P}\left(e^{tX} \geq e^{\mu t(1+\delta)}\right) \leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{\mu t(1+\delta)}} = \frac{M_X(t)}{e^{\mu t(1+\delta)}} \leq \frac{e^{\mu(e^t - 1)}}{e^{\mu t(1+\delta)}} = \left(\frac{e^{(e^t - 1)}}{e^{t(1+\delta)}}\right)^{\mu}.$$

        iii. Minimize the right-hand side in t to show that

        $$\mathbb{P}(X \geq (1 + \delta)\mu) \leq \left(e^{\delta - (1+\delta)\log(1+\delta)}\right)^{\mu}.$$

Some differentiation gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{e^{(e^t-1)}}{e^{t(1+\delta)}}\right)^{\mu} = \mu\cdot(e^t-1-\delta)\cdot\left(\frac{e^{(e^t-1)}}{e^{t(1+\delta)}}\right)^{\mu}$$

and setting that equal to 0 gives $t = \log(\delta+1)$, which the second derivative test confirms is a global minimum. Plugging this t into the right-hand side of the previous bound and rearranging a bit gives us what we want.

iv. Prove that $\delta-(1+\delta)\log(1+\delta)\leq -\delta^2/3$ for $\delta\in(0,1)$ and conclude that

$$\mathbb{P}(X\geq(1+\delta)\mu)\leq e^{-\delta^2\mu/3},$$

which is called a *Chernoff bound*. How does this compare to the kind of bound you'd get with Chebyshev?
Hint: for the first inequality, look at how the derivative of $f(x)=x-(1+x)\log(1+x)+x^2/3$ behaves on $(0,1/2)$ and $(1/2,1)$.

We want to show that the function $f(x)=x-(1+x)\log(1+x)+x^2/3$ is nonpositive for $x\in(0,1)$. Clearly $f(0)=0$, so if we can show that f is decreasing on $(0,1)$ then we're good. The first two derivatives are $f'(x)=-\log(1+x)+2x/3$ and $f''(x)=-(1+x)^{-1}+2/3$, and we note that $f''(1/2)=0$ with $f'(0), f'(1/2), f'(1)\leq 0$. So it's enough to show that f' is monotone on $(0,1/2)$ and on $(1/2,1)$, and this is easy to check using f'' (which shows that f' is decreasing on the former interval and increasing on the latter).

The Chernoff bound then follows immediately from Question 23(b)iii. To compare with Chebyshev, we'd normally have

$$\begin{aligned}
\mathbb{P}(X\geq(1+\delta)\mu) &\leq \mathbb{P}(|X-\mu|\geq\delta\mu) \\
&\leq \frac{\mathrm{Var}(X)}{\mu^2\delta^2} && \text{By Chebyshev} \\
&= \frac{\sum_{i=1}^n p_i(1-p_i)}{\mu^2\delta^2} && \text{Since the } X_i \text{ are independent and } \mathrm{Var}(X_i)=p_i(1-p_i) \\
&\leq \frac{1}{\mu\delta^2} && \text{Since } 1-p_i\leq 1 \text{ and } \mu=\textstyle\sum_i p_i
\end{aligned}$$

but with Chernoff we get $\mathbb{P}(X\geq(1+\delta)\mu)\leq e^{-\delta^2\mu/3}$, which gives an exponentially faster decay as n (and thus μ) grows. For a concrete example, take $n=100$ with $p_i=1/2$ for all i and $\delta=0.8$. Chebyshev gives the unremarkable $\mathbb{P}(X\geq(1+\delta)\mu)\leq 0.03125$, while Chernoff gives the much more impressive $\mathbb{P}(X\geq(1+\delta)\mu)\leq 0.0000233...$

24. In STA257, you may have also learned that the distribution of a random variable X is characterized by the random variable's mgf $M_X(t)$, at least when the mfg exists (a necessary condition is that $M_X(t)$ is finite when $|t|$ is arbitrarily small). Does this mean that a distribution is characterized by its integer moments? Unfortunately not. The following lognormal "family" is probably the simplest counterexample:

(a) Let

$$f(x)=\frac{1}{\sqrt{2\pi}x}\exp\left(-\frac{\log(x)^2}{2}\right),\quad x>0,$$

and for any $\varepsilon \in [-1, 1]$, let $f_\varepsilon(x) = f(x) \cdot (1 + \varepsilon \cdot \sin(2\pi \log(x)))$. Show that both $f(x)$ and $f_\varepsilon(x)$ are pdfs on $(0, \infty)$.

The first function is a special case of the second with $\varepsilon = 0$, so we can go straight for the second. With the substitution $u(x) = \log(x)$, we get

$$\begin{aligned}
\int_0^\infty f_\varepsilon(x)\,\mathrm{d}x &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{1}{x} e^{-\log(x)^2/2}\,\mathrm{d}x + \frac{\varepsilon}{\sqrt{2\pi}} \int_0^\infty \frac{1}{x} e^{-\log(x)^2/2} \sin(2\pi \log(x))\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-u^2/2}\,\mathrm{d}u + \frac{\varepsilon}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-u^2/2} \sin(2\pi u)\,\mathrm{d}u \\
&= 1 + \frac{\varepsilon}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-u^2/2} \sin(2\pi u)\,\mathrm{d}u,
\end{aligned}$$

so we just need to show that the remaining integral is 0, but that follows immediately because the integrand $e^{-u^2/2} \sin(2\pi u)$ is an odd function.

(b) Let $X \sim f$ and $Y \sim f_\varepsilon$. Show that $\mathbb{E}[X^n] = \mathbb{E}[Y^n]$ for all integers $n \geq 1$.

We have

$$\mathbb{E}[X^n] = \frac{1}{\sqrt{2\pi}} \int_0^\infty x^n \cdot \frac{1}{x} e^{-\log(x)^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_0^\infty x^{n-1} e^{-\log(x)^2/2}\,\mathrm{d}x$$

and

$$\mathbb{E}[Y^n] = \mathbb{E}[X^n] + \frac{\varepsilon}{\sqrt{2\pi}} \underbrace{\int_0^\infty x^{n-1} e^{-\log(x)^2/2} \sin(2\pi \log(x))\,\mathrm{d}x}_{=: I_n},$$

so we want to show that the remaining integral is 0. Using the same substitution $u(x) = \log(x)$ gives

$$I_n = \int_0^\infty x^{n-1} e^{-\log(x)^2/2} \sin(2\pi \log(x))\,\mathrm{d}x = \int_{-\infty}^\infty e^{nu - u^2/2} \sin(2\pi u)\,\mathrm{d}u.$$

The key now is to complete the square in the exponent. Doing so gives us

$$I_n = e^{n^2/2} \int_{-\infty}^\infty e^{-(u-n)^2/2} \sin(2\pi u)\,\mathrm{d}u = e^{n^2/2} \int_{-\infty}^\infty e^{-u^2/2} \sin(2\pi(u+n))\,\mathrm{d}u.$$

Now n is an *integer*, so that $\sin(2\pi(u+n)) = \sin(2\pi u)$ and we're dealing with the same odd integrand as in part (a), which gives $I_n = 0$.

(c) Show that $M_X(t) = \infty$ whenever $t > 0$.
Hint: the easiest way is probably to bound the integral from below by another integral that you know diverges. Use properties of the exponential function.

Fix $t > 0$. The mgf satisfies

$$\begin{aligned}
M_X(t) &= \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{tx} \cdot \frac{1}{x} e^{-\log(x)^2/2}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \exp\left(te^u - u^2/2\right)\mathrm{d}u && \text{Substituting } u(x) = \log(x) \\
&\geq \frac{1}{\sqrt{2\pi}} \int_0^\infty \exp\left(te^u - u^2/2\right)\mathrm{d}u && \text{Since the exponential function is always positive}.
\end{aligned}$$

Now, the exponential function grows faster than any polynomial, so we'll have $e^u > u^2/2t$ whenever u is sufficiently large (say whenever $u > \eta$, for some $\eta > 0$), which is the same as $te^u - u^2/2 > 0$. Therefore

$$M_X(t) \geq \frac{1}{\sqrt{2\pi}} \int_\eta^\infty \exp\left(te^u - u^2/2\right) \mathrm{d}u \geq \frac{1}{\sqrt{2\pi}} \int_\eta^\infty e^0 \, \mathrm{d}u = \infty.$$

25. Show that a continuous random variable X is symmetric around 0 (see Question 1b) if and only if X and $-X$ have the same distribution. Generalize to random variables symmetric about an arbitrary point x_0.

Let f_X and F_X be the pdf and cdf of X, respectively.

($\Rightarrow$) If X and $-X$ have the same distribution, then their cdfs satisfy $F_X(x) = F_{-X}(x)$, and the term on the right is $\mathbb{P}(-X \leq x) = \mathbb{P}(X \geq -x) = 1 - F_X(-x)$. Therefore $F_X(x) = 1 - F_X(-x)$, and differentiating gives $f_X(x) = f_X(-x)$ (i.e., X is symmetric around 0).

($\Leftarrow$) If X is symmetric around 0, then $f_X(-x) = f_X(x)$ and so

$$F_X(x) = \int_{-\infty}^x f_X(t) \, \mathrm{d}t = \int_{-\infty}^x f_X(-t) \, \mathrm{d}t = \int_{-x}^\infty f_X(u) \, \mathrm{d}u = 1 - F_X(-x),$$

while

$$F_{-X}(x) = \mathbb{P}(-X \leq x) = \mathbb{P}(X \geq -x) = 1 - F_X(-x)$$

as well. Thus $F_X = F_{-X}$, so X and $-X$ have the same distribution.

The generalization is that X is symmetric about x_0 if and only if $x_0 + X$ and $x_0 - X$ have the same distribution; the proof is essentially identical.

26. Is there a way to measure the "distance" between two probability distributions? One measure — which is not actually a metric, but still shows up all over statistics owing to its deep theoretical properties — is called the *KL divergence*. For distributions F and G supported on the same set with respective pdfs/pmfs f and g, it's defined like this:

$$D_{\mathrm{KL}}(F \,||\, G) = \mathbb{E}\left[\log\left(\frac{f(X)}{g(X)}\right)\right], \quad X \sim F.$$

(a) Calculate the KL divergence between two Poisson distributions: $D_{\mathrm{KL}}(\text{Poisson}\,(\lambda_1) \,||\, \text{Poisson}\,(\lambda_2))$.

First of all, the log-ratio of the two pmfs is

$$\log\left(\frac{f(X)}{g(X)}\right) = X \cdot \log\left(\frac{\lambda_1}{\lambda_2}\right) - (\lambda_1 - \lambda_2).$$

Therefore,

$$D_{\mathrm{KL}}(\text{Poisson}\,(\lambda_1) \,||\, \text{Poisson}\,(\lambda_2)) = \mathbb{E}[X] \cdot \log\left(\frac{\lambda_1}{\lambda_2}\right) - (\lambda_1 - \lambda_2) = \lambda_1 \cdot \log\left(\frac{\lambda_1}{\lambda_2}\right) - (\lambda_1 - \lambda_2),$$

where $X \sim \text{Poisson}\,(\lambda_1)$.

(b) Calculate the KL divergence between two exponential distributions: $D_{\mathrm{KL}}(\mathrm{Exp}\,(\lambda_1) \,||\, \mathrm{Exp}\,(\lambda_2))$.

The log-ratio of the two pdfs is

$$\log\left(\frac{f(X)}{g(X)}\right) = \log\left(\frac{\lambda_1}{\lambda_2}\right) - X \cdot (\lambda_1 - \lambda_2).$$

Therefore,

$$D_{\mathrm{KL}}(\mathrm{Exp}\,(\lambda_1) \,||\, \mathrm{Exp}\,(\lambda_2)) = \log\left(\frac{\lambda_1}{\lambda_2}\right) - \mathbb{E}\,[X] \cdot (\lambda_1 - \lambda_2) = \log\left(\frac{\lambda_1}{\lambda_2}\right) + \frac{\lambda_2}{\lambda_1} - 1.$$

where $X \sim \mathrm{Exp}\,(\lambda_1)$.

(c) Calculate the KL divergence between two normal distributions: $D_{\mathrm{KL}}(\mathcal{N}(\mu_1, \sigma_1^2) \,||\, \mathcal{N}(\mu_2, \sigma_2^2))$. *Hint*: you can do this without any integration.

This time, the log-ratio of the two pdfs is

$$\log\left(\frac{f(X)}{g(X)}\right) = \log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{(X-\mu_1)^2}{2\sigma_1^2} + \frac{(X-\mu_2)^2}{2\sigma_2^2},$$

so we want the expectation of that when $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, which is the same as

$$\log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2\sigma_1^2}\mathbb{E}\left[(X-\mu_1)^2\right] + \frac{1}{2\sigma_2^2}\mathbb{E}\left[(X-\mu_2)^2\right].$$

The first expectation is easy, because it's just $\mathrm{Var}\,(X) = \sigma_1^2$. For the second one, add and subtract μ_1 inside the squared term to get

$$\begin{aligned}\mathbb{E}\left[(X-\mu_2)^2\right] &= \mathbb{E}\left[(X-\mu_1+(\mu_1-\mu_2))^2\right] \\ &= \mathbb{E}\left[(X-\mu_1)^2\right] + \mathbb{E}\left[2(\mu_1-\mu_2)(X-\mu_1)\right] + \mathbb{E}\left[(\mu_1-\mu_2)^2\right] \\ &= \sigma_1^2 + (\mu_1-\mu_2)^2.\end{aligned}$$

Putting the pieces together gives

$$D_{\mathrm{KL}}(\mathcal{N}(\mu_1, \sigma_1^2) \,||\, \mathcal{N}(\mu_2, \sigma_2^2)) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

27. Let $X \sim F_X$ be a continuous random variable supported on $[0, b)$, for some $b > 0$. Show that

$$\mathbb{E}\,[X^n] = n\int_0^b x^{n-1} \cdot (1 - F_X(x))\,\mathrm{d}x.$$

For an extra challenge, replace b with ∞ and show the same thing (assume that $\mathbb{E}\,[X^n]$ exists to begin with). When $n = 1$ this result is called the *Darth Vader rule*, for some reason.

Let f_X be the pdf of X. For the original version,

$$\begin{aligned}\mathbb{E}\,[X^n] &= \int_0^b x^n \cdot f_X(x)\,\mathrm{d}x \\ &= x^n \cdot F_X(x)\Big|_0^b - n\int_0^b x^{n-1} \cdot F_X(x)\,\mathrm{d}x\end{aligned}$$

Using integration by parts with $u(x) = x^n$ and $\mathrm{d}v = f_X(x)\,\mathrm{d}x$

$$
\begin{aligned}
&= b^n - n\int_0^b x^{n-1}\cdot F_X(x)\,\mathrm{d}x && \text{Since } F_X(b)=1\\
&= n\int_0^b x^{n-1}\,\mathrm{d}x - n\int_0^b x^{n-1}\cdot F_X(x)\,\mathrm{d}x\\
&= n\int_0^b x^{n-1}\cdot(1-F_X(x))\,\mathrm{d}x.
\end{aligned}
$$

For the extra challenge, suppose now that X is supported on $[0,\infty)$. We can use our work above to write

$$
\begin{aligned}
\mathbb{E}[X^n] &= \lim_{b\to\infty}\left(b^n\cdot F_X(b) - n\int_0^b x^{n-1}\cdot F_X(x)\,\mathrm{d}x\right)\\
&= \lim_{b\to\infty}\left(-b^n\cdot(1-F_X(b)) + n\int_0^b x^{n-1}\cdot(1-F_X(x))\,\mathrm{d}x\right)\\
&= -\lim_{b\to\infty} b^n\cdot(1-F_X(b)) + n\int_0^\infty x^{n-1}\cdot(1-F_X(x))\,\mathrm{d}x,
\end{aligned}
$$

so the goal now is to argue that the first term is 0. First, for any $b>0$ we can write

$$
\mathbb{E}[X^n] = \mathbb{E}[X^n\cdot\mathbb{1}_{X<b}] + \mathbb{E}[X^n\cdot\mathbb{1}_{X\geq b}],
$$

which by assumption is finite. Since the left-hand side is independent of b, the same equality holds in the limit:

$$
\mathbb{E}[X^n] = \lim_{b\to\infty}\left(\mathbb{E}[X^n\cdot\mathbb{1}_{X<b}] + \mathbb{E}[X^n\cdot\mathbb{1}_{X\geq b}]\right) = \mathbb{E}[X^n] + \lim_{b\to\infty}\mathbb{E}[X^n\cdot\mathbb{1}_{X\geq b}],
$$

which forces $\lim_{b\to\infty}\mathbb{E}[X^n\cdot\mathbb{1}_{X>b}]=0$. Now, observe that for $b>0$,

$$
0\leq b^n\cdot(1-F_X(b)) = b^n\int_b^\infty f_X(x)\,\mathrm{d}x \leq \int_b^\infty x^n\cdot f_X(x)\,\mathrm{d}x = \mathbb{E}[X^n\cdot\mathbb{1}_{X\geq b}] \xrightarrow{b\to\infty} 0.
$$

By the squeeze (sandwich?) theorem, we get that $\lim_{b\to\infty} b^n\cdot(1-F_X(b))=0$, as desired.

28. Let $\mathbf{X}=(X_1,X_2)\sim\mathcal{N}_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where $\boldsymbol{\mu}=(\mu_1,\mu_2)$ and $\boldsymbol{\Sigma}=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$. Here $\mu_1,\mu_2\in\mathbb{R}$, $\sigma_1^2,\sigma_2^2>0$, and $\rho\in(-1,1)$. That is, $\mathbf{X}$ follows a bivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, which has joint pdf

$$
f_{\mathbf{X}}(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right)
$$

The goal here is to work out four things: i) the marginal distributions of X_1 and X_2, ii) the conditional distributions of $X_2\mid(X_1=x_1)$ and $X_1\mid(X_2=x_2)$, iii) the distribution of aX_1+bX_2 for $a,b\in\mathbb{R}$, and iv) the quantities $\mathrm{Cov}(X_1,X_2)$ and $\mathrm{Corr}(X_1,X_2)$. Theoretically, all of these can be found using integration and algebra alone, but that gets *very* tedious. Fortunately, there's an easier way.

(a) Let Z_1 and Z_2 be independent $\mathcal{N}(0,1)$ random variables, and let $Y_1=\mu_1+\sigma_1 Z_1$ and $Y_2=\mu_2+\sigma_2\left(\rho Z_1+\sqrt{1-\rho^2}Z_2\right)$. Prove that $(Y_1,Y_2)\overset{d}{=}(X_1,X_2)$.

The pdf of (Z_1, Z_2) is, of course,

$$f_{\mathbf{Z}}(z_1, z_2) = \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \frac{1}{2\pi} e^{-\frac{1}{2}(z_1^2+z_2^2)}.$$

Let $y_1(z_1, z_2) = \mu_1 + \sigma_1 z_1$ and $y_2(z_1, z_2) = \mu_2 + \sigma_2\left(\rho z_1 + \sqrt{1-\rho^2} z_2\right)$. Then the function $(z_1, z_2) \mapsto (y_1(z_1, z_2), y_2(z_1, z_2))$ is a smooth bijection between $\mathbb{R}$ and $\mathbb{R}$ with inverse

$$(y_1, y_2) \mapsto (z_1(y_1, y_2), z_2(y_1, y_2)) = \left(\frac{y_1 - \mu_1}{\sigma_1}, \frac{1}{\sqrt{1-\rho^2}}\left(\frac{y_2 - \mu_2}{\sigma_2} - \rho \cdot \frac{y_1 - \mu_1}{\sigma_1}\right)\right),$$

whose Jacobian has determinant

$$\det\left(\frac{\mathrm{d}(z_1, z_2)}{\mathrm{d}(y_1, y_2)}\right) = \left|\begin{bmatrix} \frac{\partial z_1}{\partial y_1} & \frac{\partial z_1}{\partial y_2} \\ \frac{\partial z_2}{\partial y_1} & \frac{\partial z_2}{\partial y_2} \end{bmatrix}\right| = \left|\begin{bmatrix} \frac{1}{\sigma_1} & 0 \\ -\frac{\rho}{\sigma_1 \cdot \sqrt{1-\rho^2}} & \frac{1}{\sigma_2 \cdot \sqrt{1-\rho^2}} \end{bmatrix}\right| = \frac{1}{\sigma_1 \sigma_2 \cdot \sqrt{1-\rho^2}},$$

so the joint pdf of (Y_1, Y_2) is

$$\begin{aligned}
&f_{\mathbf{Z}}\left(\frac{y_1 - \mu_1}{\sigma_1}, \frac{1}{\sqrt{1-\rho^2}}\left(\frac{y_2 - \mu_2}{\sigma_2} - \rho \cdot \frac{y_1 - \mu_1}{\sigma_1}\right)\right) \cdot \left|\det\left(\frac{\mathrm{d}(z_1, z_2)}{\mathrm{d}(y_1, y_2)}\right)\right| \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\left[\left(\frac{y_1 - \mu_1}{\sigma_1}\right)^2 + \frac{1}{1-\rho^2}\left(\frac{y_2 - \mu_2}{\sigma_2} - \rho \cdot \frac{y_1 - \mu_1}{\sigma_1}\right)^2\right]\right) \\
&= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{y_2 - \mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{y_1 - \mu_1}{\sigma_1}\right)\left(\frac{y_2 - \mu_2}{\sigma_2}\right)\right]\right) \\
&= f_{\mathbf{X}}(y_1, y_2).
\end{aligned}$$

Since (Y_1, Y_2) and (X_1, X_2) have the same joint pdf, we're done.

(b) Find the marginal distributions of X_1 and X_2, and then prove that X_1 and X_2 are independent if and only if $\rho = 0$.[3]

We have

$$X_1 \stackrel{d}{=} Y_1 \stackrel{d}{=} \mu_1 + \sigma_1 Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

and

$$X_2 \stackrel{d}{=} Y_2 \stackrel{d}{=} \mu_2 + \sigma_2\left(\rho Z_1 + \sqrt{1-\rho^2} Z_2\right) \sim \mathcal{N}(\mu_2, \sigma_2^2),$$

where we used the fact that Z_1 and Z_2 are *independent* (and hence $aZ_1 + bZ_2 \sim \mathcal{N}(0, a^2+b^2)$ for $a, b \in \mathbb{R}$, which is easily shown using mgfs, if you haven't already seen it).

Now, X_1 and X_2 are independent if and only if their joint pdf factorizes into the product of the marginal pdfs, and by our findings above this happens if and only if

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)\left(\frac{x_2 - \mu_2}{\sigma_2}\right)\right]\right)$$

[3] In other words: if a pair of normal random variables jointly follows a bivariate normal distribution, then the (normally distributed) marginals are independent if and only if they're uncorrelated. Unfortunately, students tend to forget about the qualifier at the start of that statement, resulting in the extremely common and extremely incorrect misconception that "two normal random variables are independent if and only if they're uncorrelated." *Please never say this.*

$$= \left[\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right)\right]\cdot\left[\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)\right]$$

for all $x_1, x_2 \in \mathbb{R}$, which (by inspection) happens if and only if $\rho = 0$.

(c) Find the conditional distributions of $X_2 \mid (X_1 = x_1)$ and $X_1 \mid (X_2 = x_2)$.
Hint: after finding the first one, argue how the second follows immediately by symmetry.

First of all, using the equivalent joint distribution above, this is the same thing the conditional distribution of $\mu_2 + \sigma_2\left(\rho Z_1 + \sqrt{1-\rho^2}Z_2\right)$ given $Z_1 = (x_1-\mu_1)/\sigma_1$, which is the distribution of

$$\mu_2 + \sigma_2 \cdot \left(\rho\frac{x_1-\mu_1}{\sigma_1} + \sqrt{1-\rho^2}Z_2\right) = \left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\right) + \sqrt{1-\rho^2}\sigma_1 Z_2$$

which, using the fact that $a + bZ_1 \sim \mathcal{N}(a, b^2)$ for $a, b \in \mathbb{R}$, is

$$\mathcal{N}\left(\mu_2 + \rho\sigma_2\frac{x_1-\mu_1}{\sigma_1}, (1-\rho^2)\sigma_2^2\right).$$

Since $f_{\mathbf{X}}(x_1, x_2)$ is symmetric in $(x_1-\mu_1)/\sigma_1$ and $(x_2-\mu_2)/\sigma_2$, all we need to do to get the distribution of $X_1 \mid (X_2 = x_2)$ is swap μ_1 with μ_2 and σ_1 with σ_2 above, which gives us

$$X_1 \mid (X_2 = x_2) \sim \mathcal{N}\left(\mu_1 + \rho\sigma_1\frac{x_2-\mu_2}{\sigma_2}, (1-\rho^2)\sigma_1^2\right).$$

(d) Let $a, b \in \mathbb{R}$. Find the distribution of $aX_1 + bX_2$.

Again, using the equivalent joint distribution above, this is the same thing as

$$a\left(\mu_1 + \sigma_1 Z_1\right) + b\left(\mu_2 + \sigma_2\left(\rho Z_1 + \sqrt{1-\rho^2}Z_2\right)\right) = (a\mu_1 + b\mu_2) + (a\sigma_1 + b\sigma_2\rho)Z_1 + b\sqrt{1-\rho^2}\sigma_2 Z_2,$$

which gives

$$aX_1 + bX_2 \sim \mathcal{N}\left(a\mu_1 + b\mu_2, a^2\sigma_1^2 + 2ab\rho\sigma_1\sigma_2 + b^2\sigma_2^2\right).$$

(e) Find $\mathrm{Cov}(X_1, X_2)$ and $\mathrm{Corr}(X_1, X_2)$.

The first one is

$$\begin{aligned}
\mathrm{Cov}(X_1, X_2) &= \mathrm{Cov}\left(\mu_1 + \sigma_1 Z_1, \mu_2 + \sigma_2\left(\rho Z_1 + \sqrt{1-\rho^2}Z_2\right)\right) \\
&= \mathrm{Cov}\left(\sigma_1 Z_1, \sigma_2\rho Z_1 + \sqrt{1-\rho^2}\sigma_2 Z_2\right) \\
&= \mathrm{Cov}\left(\sigma_1 Z_1, \sigma_2\rho Z_1\right) + \mathrm{Cov}\left(\sigma_1 Z_1, \sqrt{1-\rho^2}\sigma_2 Z_2\right) \\
&= \mathrm{Cov}\left(\sigma_1 Z_1, \sigma_2\rho Z_1\right) \\
&= \rho\sigma_1\sigma_2\mathrm{Var}\left(Z_1\right) \\
&= \rho\sigma_1\sigma_2
\end{aligned}$$

and the second is

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}\left(X_1\right)\cdot\mathrm{Var}\left(X_2\right)}} = \frac{\rho\sigma_1\sigma_2}{\sqrt{\sigma_1^2\sigma_2^2}} = \rho.$$

*29. Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and let $f(x, A) = \mathbb{P}(X \geq x \mid X \in A)$, where $A \subseteq \mathbb{R}$ is some set. Letting $Z \sim \mathcal{N}(0, 1)$ and using the standard normal cdf $\Phi(\cdot)$ if need be, compute the following:

(a) $f(\mu, (-\infty, \mu])$

This is

$$\mathbb{P}(X \geq \mu \mid X \leq \mu) = \frac{\overbrace{\mathbb{P}(X = \mu)}^{=0}}{\mathbb{P}(X \leq \mu)} = 0.$$

(b) $f(\mu, \mathbb{R})$

This is

$$\mathbb{P}(X \geq \mu \mid X \in \mathbb{R}) = \frac{\mathbb{P}(X \geq \mu)}{\underbrace{\mathbb{P}(X \in \mathbb{R})}_{=1}} = \mathbb{P}\left(\frac{X - \mu}{\sigma} \geq 0\right) = \mathbb{P}(Z \geq 0) = \frac{1}{2}.$$

(c) $f(-\mu, [-\mu, \infty))$

This is

$$\mathbb{P}(X \geq -\mu \mid X \geq -\mu) = \frac{\mathbb{P}(X \geq -\mu)}{\mathbb{P}(X \geq -\mu)} = 1.$$

(d) $f(\mu, \mathbb{R} \setminus (-\mu, \mu))$

This is

$$\mathbb{P}(X \geq \mu \mid X \leq -\mu \vee X \geq \mu) = \frac{\mathbb{P}(X \geq \mu)}{\mathbb{P}(X \leq -\mu) + \mathbb{P}(X \geq \mu)} = \frac{1/2}{\Phi(-2\mu/\sigma) + 1/2}$$

because $\mathbb{P}(X \geq \mu) = 1/2$ and $\mathbb{P}(X \leq -\mu) = \mathbb{P}((X - \mu)/\sigma \leq -2\mu/\sigma) = \Phi(-2\mu/\sigma)$.

(e) $f(\mu + k\sigma, [\mu + j\sigma, \infty))$, where $k, j \in \mathbb{N}$

This is

$$\mathbb{P}(X \geq \mu + k\sigma \mid X \geq \mu + j\sigma) = \frac{\mathbb{P}(X \geq \max\{\mu + k\sigma, \mu + j\sigma\})}{\mathbb{P}(X \geq \mu + j\sigma)} = \frac{\mathbb{P}(Z \geq \max\{k, j\})}{\mathbb{P}(Z \geq j)} = \frac{1 - \Phi(\max\{k, j$$

(f) $f(Y, \mathbb{R})$, where $Y \sim \mathcal{N}(\mu, \sigma^2)$ is independent of X

Since X and Y are independent, $X - Y \sim \mathcal{N}(0, 2\sigma^2)$, so[4]

$$\mathbb{P}(X \geq Y \mid X \in \mathbb{R}) = \mathbb{P}(X - Y \geq 0) = \mathbb{P}(Z \geq 0) = \frac{1}{2}.$$

[4] Actually, $\mathbb{P}(X \geq Y) = 1/2$ for any continuous, independent, and identically distributed random variables X and Y. To see this, observe that we must have $1 = \mathbb{P}(X > Y) + \mathbb{P}(X \leq Y) = \mathbb{P}(X \geq Y) + \mathbb{P}(X \leq Y)$, and since (X, Y) and (Y, X) clearly have the same joint distributions, those two probabilities on the right must be equal. Intuitively, either X exceeds Y or Y exceeds X, and since X and Y independent and follow the same distribution, neither one of those events should have a higher/lower probability than the other.

(g) $f(Y + \sqrt{3}\sigma, \mathbb{R})$, where $(X, Y) \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (\mu, \mu)$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & -\sigma^2/2 \\ -\sigma^2/2 & \sigma^2 \end{bmatrix}$

From Question 28d, we see that $X - Y \sim \mathcal{N}(0, 3\sigma^2)$, which gives

$$\mathbb{P}\left(X \geq Y + \sqrt{3}\sigma \mid X \in \mathbb{R}\right) = \mathbb{P}\left(X - Y \geq \sqrt{3}\sigma\right) = \mathbb{P}\left(Z \geq 1\right) = 1 - \Phi(1).$$

(h) $\mathbb{E}\left[f(\mu, (-\infty, Y])\right]$, where $Y \sim \mathcal{N}(\mu, \sigma^2)$ is independent of X

Conditioning on Y, the random variable inside the expectation is

$$\begin{aligned}
\mathbb{P}\left(X \geq \mu \mid X \leq Y\right) &= \frac{\mathbb{P}\left(\mu \leq X \leq Y\right)}{\mathbb{P}\left(X \leq Y\right)} \cdot \mathbb{1}_{\mu \leq Y} \\
&= \frac{\mathbb{P}\left(0 \leq Z \leq (Y - \mu)/\sigma\right)}{\mathbb{P}\left(Z \leq (Y - \mu)/\sigma\right)} \cdot \mathbb{1}_{0 \leq (Y-\mu)/\sigma} \\
&= \frac{\Phi((Y - \mu)/\sigma) - \Phi(0)}{\Phi((Y - \mu)/\sigma)} \cdot \mathbb{1}_{0 \leq (Y-\mu)/\sigma} \\
&\overset{d}{=} \left(1 - \frac{\Phi(0)}{\Phi(Z)}\right) \cdot \mathbb{1}_{0 \leq Z} \\
&= \mathbb{1}_{0 \leq Z} - \frac{\Phi(0)}{\Phi(Z)} \cdot \mathbb{1}_{0 \leq Z}
\end{aligned}$$

The expectation of that is

$$\mathbb{P}\left(0 \leq Z\right) - \mathbb{E}\left[\frac{\Phi(0)}{\Phi(Z)} \cdot \mathbb{1}_{0 \leq Z}\right] = 1 - \Phi(0) - \Phi(0) \cdot \int_0^\infty \frac{1}{\Phi(z)} \cdot \frac{e^{-z^2/2}}{\sqrt{2\pi}} \, \mathrm{d}z.$$

Making the substitution $u(z) = \Phi(z)$ with $\mathrm{d}u = e^{-z^2/2}/\sqrt{2\pi}\,\mathrm{d}z$ turns the integral into

$$\int_{\Phi(0)}^{\Phi(\infty)} \frac{1}{u} \, \mathrm{d}u = \log\left(u\right)\Big|_{1/2}^{1} = \log\left(2\right),$$

so in the end, the expectation is

$$1 - \frac{1}{2} - \frac{1}{2}\log\left(2\right) = \frac{1 - \log\left(2\right)}{2}.$$

30. Fix $q > 0$. Find a continuous random variable X and a discrete random variable Y such that $\mathbb{E}\left[X^q\right] = \mathbb{E}\left[Y^q\right] = \infty$, but $\mathbb{E}\left[X^p\right], \mathbb{E}\left[Y^p\right] < \infty$ for all $0 \leq p < q$.

There are many ways to do this; I find that the easiest is to start with an integral/sum that we know to converge or diverge based on a certain parameter, and then tweak the integrand/summand so that it looks like a familiar pdf/pmf times some function. In the continuous case, for example, it's easy to show (using integration by parts) that

$$\int_0^\infty x^{-\alpha} \cdot e^{-x} \, \mathrm{d}x \begin{cases} = \infty & \text{if } \alpha \geq 1 \\ < \infty & \text{if } \alpha < 1 \end{cases}$$

which means that if $p < q$, then

$$\int_0^\infty \left(x^{-(1/q)}\right)^q \cdot e^{-x} \, \mathrm{d}x = \infty \qquad \text{but} \qquad \int_0^\infty \left(x^{-(1/q)}\right)^p \cdot e^{-x} \, \mathrm{d}x < \infty.$$

So taking $X = V^{-1/q}$ where $V \sim \text{Exp}(1)$ does the trick.

For the discrete version, we can apply the same idea to a p-series, since we know

$$\sum_{n=1}^{\infty} \frac{1}{n^\alpha} \begin{cases} = \infty & \text{if } \alpha \leq 1 \\ < \infty & \text{if } \alpha > 1 \end{cases}.$$

So, for example, let W be such that $\mathbb{P}(W = n) \propto 1/n^2$ for $n = 1, 2, \ldots$ and take $Y = W^{1/q}$.

31. Let X be a random variable with a finite second moment. Prove *Cantelli's inequality*: for any $\lambda > 0$, we have

$$\mathbb{P}(X - \mathbb{E}[X] \geq \lambda) \leq \frac{\text{Var}(X)}{\text{Var}(X) + \lambda^2}.$$

Hint: Upper bound the left-hand side by $\mathbb{P}\left((X - \mathbb{E}[X] + x)^2 \geq (\lambda + x)^2\right)$ for any $x \in \mathbb{R}$. Then apply Markov's inequality and optimize over x.

Following the hint, we see that for any $x \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}(X - \mathbb{E}[X] \geq \lambda) &= \mathbb{P}(X - \mathbb{E}[X] + x \geq \lambda + x) \\ &\leq \mathbb{P}\left((X - \mathbb{E}[X] + x)^2 \geq (\lambda + x)^2\right) \\ &\leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X] + x)^2\right]}{(\lambda + x)^2} && \text{By Markov's inequality} \\ &= \frac{\text{Var}(X) + x^2}{(\lambda + x)^2} && \text{After expanding the square as } (X-\mathbb{E}[X])^2+2x(X-\mathbb{E}[X])+x^2. \text{ and taking expectations} \end{aligned}$$

Now, the derivative of the last term with respect to x is

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\text{Var}(X) + x^2}{(\lambda + x)^2}\right) = \frac{2x(\lambda + x) - 2(\text{Var}(X) + x^2)}{(\lambda + x)^3}$$

and setting this to 0 gives $x = \text{Var}(X)/\lambda$, which the second derivative test confirms is a minimum. Plugging this into our upper bound gives us what we're looking for.

*32. The *Cauchy-Schwarz inequality* is one of the most ubiquitous inequalities in math; there's a good chance you've seen it before in one setting or another. Here's a version that we'll need in our course, which is often called the *covariance inequality*: for any random variables X, Y with finite second moments,

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X) \cdot \text{Var}(Y)}, \tag{1}$$

where equality holds if and only if X is a certain linear function of Y (with probability 1). Let's prove it! To be proper, we'll declare right here that all statements about X and Y in this question implicitly hold with probability 1.[5]

(a) Prove the result when either $\text{Var}(X) = 0$ or $\text{Var}(Y) = 0$. With that taken care of, assume going forward (without loss of generality) that $\text{Var}(Y) > 0$.

Suppose that $\text{Var}(X) = 0$. Obviously the right-hand side of (1) is 0, so we want to show that the left-hand side is also 0. But if $\text{Var}(X) = 0$, then X is constant;[6] say $X = x$ for some $x \in \mathbb{R}$. Therefore

$$|\text{Cov}(X, Y)| = |\mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]| = |x \cdot \mathbb{E}[Y] - x \cdot \mathbb{E}[Y]| = 0.$$

[5] In other words, if we say something like $X = Y$, we really mean that $\mathbb{P}(X = Y) = 1$. It's okay to ignore this technicality here because this question is about expectations, and expectations don't care about events of probability 0.

[6] If you haven't seen this in STA257, here's one way to show it: since $0 = \text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ and the thing inside the expectation is always non-negative, we must have $(X - \mathbb{E}[X])^2 = 0$, which is the same as $X = \mathbb{E}[X]$. In other words, X must be equal to its own expectation.

(b) Show that the function $f(t) = \mathbb{E}\left[(X - tY)^2\right]$ is quadratic in t, and explain why it must have at most one real root.

Expand the square to get

$$f(t) = \mathbb{E}\left[X^2\right] - 2t \cdot \mathbb{E}\left[XY\right] + t^2 \cdot \mathbb{E}\left[Y^2\right],$$

which is indeed a quadratic function of t. Furthermore, it's non-negative because it's the expectation of the non-negative thing $(X - tY)^2$. A non-negative quadratic function looks like a convex (or "concave up") parabola that never dips below the x-axis. If the parabola just touches the x-axis, then it has exactly one real root; otherwise, it has none at all.

(c) Think back to the quadratic formula and use the last fact to obtain

$$|\mathbb{E}\left[XY\right]| \leq \sqrt{\mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]}. \tag{2}$$

The quadratic formula says that the roots of the quadratic $at^2 + bt + c$ are given by

$$t_\pm = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

and whether the $t_\pm$ are real or complex depends solely on the discriminant $b^2 - 4ac$. If the quadratic has at most one real root, then exactly one of two things are true: either it has *no* real roots (in which case $b^2 - 4ac < 0$), or it has a real root with multiplicity 2 (i.e., a double root, in which case $b^2 - 4ac = 0$). So the statement "$at^2 + bt + c$ has at most one real root" is the same thing as "$b^2 - 4ac \leq 0$". Substituting $a = \mathbb{E}\left[Y^2\right]$, $b = -2\mathbb{E}\left[XY\right]$, and $c = \mathbb{E}\left[X^2\right]$ means that $4\mathbb{E}\left[XY\right]^2 - 4\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right] \leq 0$, which is equivalent to (2).

(d) Show that equality in (2) holds if and only if $X = t^*Y$, where $t^* = \mathbb{E}\left[XY\right]/\mathbb{E}\left[Y^2\right]$.

($\Rightarrow$) If $X = t^*Y$, then

$$|\mathbb{E}\left[XY\right]| = \left|t^* \cdot \mathbb{E}\left[Y^2\right]\right| = |t^*|\sqrt{\mathbb{E}\left[Y^2\right] \cdot \mathbb{E}\left[Y^2\right]} = \sqrt{\mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]},$$

so equality holds in (2).

($\Leftarrow$) In the derivation of (2), we saw that equality holds when the quadratic $f(t)$ has a double root which, by the quadratic formula above, is given by

$$t^* = -\frac{b}{2a} = \frac{\mathbb{E}\left[XY\right]}{\mathbb{E}\left[Y^2\right]}.$$

Then $0 = f(t^*) = \mathbb{E}\left[(X - t^*Y)^2\right]$, which gives us $X = t^*Y$.

(e) Obtain (1) by replacing X and Y in (2) with $X - \mathbb{E}\left[X\right]$ and $Y - \mathbb{E}\left[Y\right]$, respectively. Exactly when does equality hold?

(1) falls out immediately after performing the replacement. From (d), we know that equality holds if and only if $X - \mathbb{E}\left[X\right] = t^*(Y - \mathbb{E}\left[Y\right])$, where

$$t^* = \frac{\mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right]}{\mathbb{E}\left[(Y - \mathbb{E}\left[Y\right])^2\right]} = \frac{\mathrm{Cov}\left(X, Y\right)}{\mathrm{Var}\left(Y\right)}.$$

Equivalently, equality holds in (1) if and only if

$$X = t^*Y + s^*,$$

where

$$s^* = \mathbb{E}[X] - \mathbb{E}[Y] \cdot \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}.$$

33. Prove the *Paley-Zygmund inequality*: if X is a non-negative random variable with a finite second moment, then for any $\lambda \in [0,1]$,

$$\mathbb{P}(X > \lambda \cdot \mathbb{E}[X]) \geq (1-\lambda)^2 \cdot \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}.$$

Hint: start by writing $X = X \cdot \mathbb{1}_{X \leq \lambda \cdot \mathbb{E}[X]} + X \cdot \mathbb{1}_{X > \lambda \cdot \mathbb{E}[X]}$, take expectations, and use (2) somewhere.

Following the hint, we have

$$\mathbb{E}[X] = \mathbb{E}\left[X \cdot \mathbb{1}_{X \leq \lambda \cdot \mathbb{E}[X]}\right] + \mathbb{E}\left[X \cdot \mathbb{1}_{X > \lambda \cdot \mathbb{E}[X]}\right].$$

Inside the first expectation on the right, we have that $X \cdot \mathbb{1}_{X \leq \lambda \cdot \mathbb{E}[X]} \leq \lambda \cdot \mathbb{E}[X]$ (because if $X > \lambda \cdot \mathbb{E}[X]$ then the indicator function is 0), so by taking expectations we get $\mathbb{E}\left[X \cdot \mathbb{1}_{X \leq \lambda \cdot \mathbb{E}[X]}\right] \leq \mathbb{E}[\lambda \cdot \mathbb{E}[X]] = \lambda \cdot \mathbb{E}[X]$. On the other hand, applying (2) to the second expectation gives

$$\begin{aligned}
\mathbb{E}\left[X \cdot \mathbb{1}_{X > \lambda \cdot \mathbb{E}[X]}\right] &\leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}\left[\mathbb{1}^2_{X > \lambda \cdot \mathbb{E}[X]}\right]} \\
&= \sqrt{\mathbb{E}[X^2] \cdot \mathbb{E}\left[\mathbb{1}_{X > \lambda \cdot \mathbb{E}[X]}\right]} \\
&= \sqrt{\mathbb{E}[X^2] \cdot \mathbb{P}(X > \lambda \cdot \mathbb{E}[X])}.
\end{aligned}$$

Therefore,

$$\mathbb{E}[X] \leq \lambda \cdot \mathbb{E}[X] + \sqrt{\mathbb{E}[X^2] \cdot \mathbb{P}(X > \lambda \cdot \mathbb{E}[X])}.$$

Equivalently,

$$(1-\lambda) \cdot \mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2] \cdot \mathbb{P}(X > \lambda \cdot \mathbb{E}[X])}.$$

Squaring and rearranging gives us what we want.

34. Let X be a random variable taking values in the non-negative integers (assume this for all random variables in this question) whose moments exist. The *probability generating function (pgf)* of X is the function $G_X(t) = \mathbb{E}\left[t^X\right] = \sum_{j=0}^{\infty} \mathbb{P}(X=j) \cdot t^j$.

(a) Show that $\mathbb{E}[X] = G'_X(1)$ and $\mathrm{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$

We have

$$G'_X(t) = \sum_{j=0}^{\infty} j \cdot \mathbb{P}(X=j) \cdot t^{j-1} \qquad \text{so that} \qquad G'_X(1) = \sum_{j=0}^{\infty} j \cdot \mathbb{P}(X=j) = \mathbb{E}[X]$$

and

$$G''_X(t) = \sum_{j=0}^{\infty} j \cdot (j-1) \cdot \mathbb{P}(X=j) \cdot t^{j-2} = \sum_{j=0}^{\infty} j^2 \cdot \mathbb{P}(X=j) \cdot t^{j-2} - \sum_{j=0}^{\infty} j \cdot \mathbb{P}(X=j) \cdot t^{j-2}$$

so that

$$G''_X(1) = \sum_{j=0}^{\infty} j^2 \cdot \mathbb{P}(X = j) - \sum_{j=0}^{\infty} j \cdot \mathbb{P}(X = j) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]$$

and

$$\mathrm{Var}\left(X\right) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 = G''_X(1) + +G'_X(1) - \left[G'_X(1)\right]^2.$$

(b) If $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with pgf $G_X(t)$, and N is another random variable independent of the X_i's with pgf $G_N(t)$, show that the pgf of $Y = \sum_{j=1}^{N} X_j$ is $G_N(G_X(t))$.

We have

$$\begin{aligned}
G_Y(t) &= \mathbb{E}\left[t^{\sum_{j=1}^{N} X_j}\right] \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left[t^{\sum_{j=1}^{n} X_j}\right] \cdot \mathbb{P}\left(N = n\right) && \text{By the law of total expectation (i.e., the "tower rule")} \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left[t^{X_1}\right]^n \cdot \mathbb{P}\left(N = n\right) && \text{Since the } X_i \text{ are independent and identically distributed} \\
&= \sum_{n=0}^{\infty} G_X(t)^n \cdot \mathbb{P}\left(N = n\right) \\
&= G_N(G_X(t)).
\end{aligned}$$

(c) Find the pgfs of the Binomial(k, p), the Poisson(λ), and the Geometric(p) distributions. If there are infinite series involved, assume whatever values of t you need to make them converge.

For $X \sim$ Binomial(k, p):

$$\begin{aligned}
G_X(t) &= \sum_{j=0}^{k} \mathbb{P}\left(X = j\right) \cdot t^j && \text{Since } X \text{ can only take values in } \{0, \ldots, k\} \\
&= \sum_{j=0}^{k} \binom{k}{j} \cdot p^j \cdot (1-p)^{k-j} \cdot t^j \\
&= \sum_{j=0}^{k} \binom{k}{j} \cdot (tp)^j \cdot (1-p)^{k-j} \\
&= (tp + (1-p))^k. && \text{By the binomial theorem}
\end{aligned}$$

For $Y \sim$ Poisson$\left(\lambda\right)$:

$$G_Y(t) = \sum_{j=0}^{\infty} \mathbb{P}\left(Y = j\right) \cdot t^j = \sum_{j=0}^{\infty} \frac{\lambda^j \cdot e^{-\lambda}}{j!} \cdot t^j = \sum_{j=0}^{\infty} \frac{(t\lambda)^j \cdot e^{-\lambda}}{j!} = e^{\lambda \cdot (t-1)}.$$

For $Z \sim$ Geometric(p), using the version of the distribution supported on $\{1, 2, \ldots\}$:

$$G_Z(t) = \sum_{j=0}^{\infty} \mathbb{P}\left(Z = j\right) \cdot t^j \qquad \text{Since } Z \text{ cannot take on the value 0}$$

$$= \sum_{j=1}^{\infty} (1-p)^{j-1} p \cdot t^j$$

$$= tp \sum_{k=0}^{\infty} (t(1-p))^k$$

Substituting $k = j - 1$

$$= \frac{tp}{1 - t(1-p)}.$$

Summing the geometric series, provided the series converges (i.e., when $|t| < 1/(1-p)$)